Geometric Algorithms for Protein Structure Determination Using Measurements From Nuclear Magnetic Resonance Spectroscopy

by

Jeffrey W. Martin

Department of Computer Science
Duke University

Date: _____

Approved:

_____
Bruce R. Donald, Supervisor

_____
Pankaj K. Agarwal

_____
Alexander J. Hartemink

_____
Pei Zhou

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Computer Science
in the Graduate School of Duke University
2014

Abstract

Geometric Algorithms for Protein Structure Determination Using Measurements From Nuclear Magnetic Resonance Spectroscopy

by

Jeffrey W. Martin

Department of Computer Science
Duke University

Date: _____

Approved:

_____
Bruce R. Donald, Supervisor

_____
Pankaj K. Agarwal

_____
Alexander J. Hartemink

_____
Pei Zhou

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science in the Graduate School of Duke University
2014

Abstract

In an environment such as a cell, the three-dimensional structure of a protein entirely determines its function. Hence, to understand the mechanics of biochemical processes necessary to sustain life, it is crucial to study the structures of proteins at atomic detail. When life is threatened by viral and bacterial pathogens, structural characterization of the proteins at play yields insights about possible treatments and therapeutics. Measurements from nuclear magnetic resonance spectroscopy (NMR) reveal information about the structures of proteins, but building accurate atomic-resolution models from such measurements is an arduous task. The ambiguity and uncertainty of these measurements, and the challenges of obtaining a sufficient number of measurements to uniquely describe a structure, contribute to the difficulty of protein structure determination by NMR.

The current widely-used computational methods using NMR measurements for structure determination primarily rely on various incarnations of stochastic optimization. These techniques have been used to determine protein structures of excellent quality, but in the long term, the reliability of these techniques is dubious (and in cases, demonstrably inadequate), especially as we attempt to solve increasingly difficult structures. Stochastic optimization, due to its random nature, may not always report the best solution. Other superior solutions may lie concealed in the landscape of the objective function and remain undiscovered. We therefore seek computational methods for structure determination that are imbued with guarantees about solution

quality. In this dissertation, we present methods for protein structure determination by NMR that are able to guarantee structural solutions quantitatively agree with experimental measurements. Although the trade-off for guaranteeing completeness of algorithms for structure determination is often an exponential running time (assuming $P \neq NP$), for some methods, we remarkably obtained polynomial running times in addition to guarantees of completeness.

To Mom, with love.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my advisor, Prof. Bruce Donald, for his guidance over the years. As his student, not only have I learned how to perform and present research, but this experience has also taught me how to thrive in the competitive academic environment, for which I am very grateful. Bruce's expert skill at building and maintaining productive collaborations meant there has never been a shortage of opportunities for great research.

I am also grateful to my dissertation committee members. I thank Prof. Pankaj Agarwal for always being willing to share his expertise about geometry. I thank Prof. Alexander Hartemink for always encouraging us to approach problems from a statistical point of view. I thank Prof. Pei Zhou for being an excellent collaborator and co-author who is never short on profoundly useful suggestions that have often led to exciting new research ideas.

I would like to thank Prof. Leonard Spicer for being an excellent collaborator. His advice on research and the hurdles of academic life have been invaluable. I also thank Dr. Patrick Reardon for being both a collaborator and a friend. Finally, all those years of hard work on the MPER project paid off! I also thank Prof. Terrence Oas and Yang Qi for being great collaborators that are full of interesting mathematical problems in need of solutions.

My lab mates, both current and former, have been great comrades. In particular, I am grateful to Dr. Anthony Yan for always being genuinely excited to explain

hard math. Many of my research projects have been founded on ideas he pioneered. I thank Prof. Jianyang (Michael) Zeng, Dr. Chittaranjan Tripathy, and Dr. Anna Yershova for their exciting discussions about NMR, math, geometry, proteins, and structural biology. I also thank Kyle Roberts, Pablo Gainza, and John MacMaster for always being engaged in discussion. I thank my office mate, Jonathan Jou, for witty banter. I also thank Michael Hemmer for helping to analyze the geometry of super difficult curves on the sphere.

I thank the Duke computer science lunch group, Dr. Thomas Mølhave, Dr. Troels Sørensen, Neil Lebeck, David Becker, and Matt Saylor, whose membership has rotated over the years, but we have always eaten in good company.

My mother, to whom this dissertation is dedicated, deserves special thanks. She has always encouraged me to pursue academic studies and has gone to great lengths to make sure I had ample opportunities to do so. Thanks mom!

Thanks to Mike Gratton for saving me tons of time preparing this dissertation by graciously making his LaTeX template available for all to use. Although I never met him personally, Mel Gratton was a very energetic and entertaining acquaintance.

I thank all my Durham, Chapel Hill, and Raleigh friends who are all surely too numerous to name. I will always have freshly brewed beer on tap for Halcy, Liz, Firas, Hero, Eduardo, Ned, Stephen (classic), John, and Alex. The shenanigans perpetrated by Viki, Brittany, Dave, Josh, Linda, Jeff, Bei, Miles, Laura, Steven (new), Ben, Drew, Ryland, Paul, Shadoe, Rachael, Michael, and others gave balance to a life beset by deadlines. Finally, I thank my best friend, Ben, and his lady Shannon, for traveling at great expense to celebrate my defense.

1

Introduction

This dissertation is organized into six chapters. Chapter 1 (this chapter) presents a brief introduction to structural biology and NMR. The following chapters present several algorithms for analysis of NMR data and protein structure determination – particularly the structure determination of difficult symmetric proteins composed of identical subunits, termed *homo-oligomers*. An algorithm for the determination of symmetric homo-oligomeric proteins and our DISCO software implementation is presented in Chapter 2. Chapter 3 describes our work applying DISCO to help determine the structure of an HIV protein, the membrane proximal external region of gp41 (MPER).

A conflict in the literature over the structure of a difficult membrane-associated symmetric homo-oligomeric protein, Diacylglycerol kinase from *Escherichia coli*, triggered a re-analysis of generally accepted practice for structure determination by NMR. We present in Chapter 4 a case study where a widely-used NMR structure determination protocol produced a structure that disagrees with a completely separate structure determination protocol that relied on fundamentally different experimental methods. We conclude that the problem is rooted in the algorithms used in the NMR

structure determination protocol, and present our solution.

Chapter 5 discusses the usage of orientational information from NMR experiments for structure determination. Traditional and widely-used protocols for NMR structure determination have difficulty using orientational information to define the overall fold of the protein. Instead, these protocols often use inter-atomic distance information to define the overall protein fold, and the orientational information is used in a later refinement step after much of the structure has already been determined. Previous work in the Donald lab created a framework that uses largely orientational information to define the overall fold of a protein. Chapter 5 presents an algorithmic module that fits into this framework. The goal of this module is to determine structures of protein fragments using orientational information alone which will remove the last dependence on distance information from the framework.

Finally, Chapter 6 presents LibProtNmr, an open-source software library written in Java that implements a vast array of protein structure manipulation and NMR data analysis techniques. This library provides an application programming interface (API) to perform many low-level tasks needed by structural biology algorithms and is the culmination of several years of software development.

Some chapters of this dissertation are presented in peer-reviewed publications. Chapter 2 is based on the following publications.

J. W. Martin, A. K. Yan, C. Bailey-Kellogg, P. Zhou, and B. R. Donald. Protein Science, 2011. 20(6):970–985.

J. W. Martin, A. K. Yan, C. Bailey-Kellogg, P. Zhou, and B. R. Donald. Journal of Computational Biology, 2011. 18(11):1507–1523.

Chapter 3 is based on the following publication.

P. N. Reardon, H. Sage, S. M. Dennison, J. W. Martin, B. R. Donald, S. M. Alam, B. F. Haynes, and L. D. Spicer. Proceedings of the National Academy of Sciences,

2014. 111(4):1391–1396.

Chapter 4 is based on a manuscript that is currently in submission. Chapters 5 and 6 are currently unpublished.

1.1 Structural biology

Structural characterization of proteins yields insight into their biological functions, which has become increasingly important for understanding the biochemical basis of human disease. Pathogenic organisms and viruses rely heavily on proteins to perform molecular tasks to infect their hosts and propagate their deleterious effects. For instance, the human immunodeficiency virus (HIV) causes acquired immunodeficiency syndrome (AIDS) in humans which leads to a drastically reduced effectiveness of the immune system, hence leaving the affected individual vulnerable to additional (and often life-threatening) infections. HIV-1 (the more virulent type of HIV (Gilbert et al., 2003)) employs a host of different proteins to perform various tasks throughout its replication cycle. The HIV-1 genome encodes an envelope of proteins to surround itself. One of these proteins, gp160, is inactive until targeted by a cellular protease which cleaves it into two separate proteins: gp120 and gp41 (McCune et al., 1988; Chan et al., 1997). These two proteins form a complex called the viral spike which is believed to participate in one of the mechanisms responsible for viral infection of cells (Kwong et al., 1998). The viral spike is thought to mediate membrane fusion which merges the contents of the viral envelope with the host cell, hence allowing the viral DNA remodeling proteins access to the genome of the cell. A virally-encoded reverse transcriptase converts the relatively short RNA genome of the virus into DNA (Kohlstaedt et al., 1992). This DNA is later included into the genome of the host by integrase, another viral protein (Bushman et al., 1993). Once embedded into the genome of the host cell, the virus can be in a productive state, causing the host cell to destroy itself producing more copies of the virus, or the virus

can lie dormant in the genome until subsequent expression, hence completely evading antiretroviral therapies (Margolis and Archin, 2006). Understanding the functions of proteins employed by pathogens, such as HIV-1, will eventually reveal new therapies, treatments, and vaccines for these diseases.

To understand the mechanisms of disease, like viral infection, it is important to study the functions of the proteins involved. In the cellular environment, each protein begins its life as an elongated chain of amino acids and then quickly collapses into a specific three-dimensional structure (See Figure 1.1 for common protein structural features). The particular shape adopted by the protein in its three-dimensional conformation dictates what function it will play in the cell. Therefore, one must study the structures of proteins to fully understand their functions and how they are performed. In particular, the individual locations of the constituent atoms of the protein are extremely important to help describe how the protein can participate in reactions with partner molecules. For this reason, structural models of proteins at *atomic-resolution* (meaning, the relative location of each atom is known) are the most useful models. Although it is possible to build models of proteins where only the rough shape is known, these models generally lack sufficient detail to explain how observed reactions are performed, and hence usually serve as scaffolds for further structural characterization using atomic-resolution methodology (DiMaio et al., 2009).

In many cases, it is possible to observe the function of a protein, but the details of how the function is actually performed is unknown. For instance, the general mechanism of cell infection by an HIV-1 virion is fairly well-understood (Tilton and Doms, 2010), but the molecular details of how the viral membrane is able to fuse to the outer membrane of the cell are still under investigation (Buzon et al., 2010; Tran et al., 2012). In these cases, building three-dimensional atomic resolution structures of the proteins involved (in particular, gp120 and gp41) yields additional

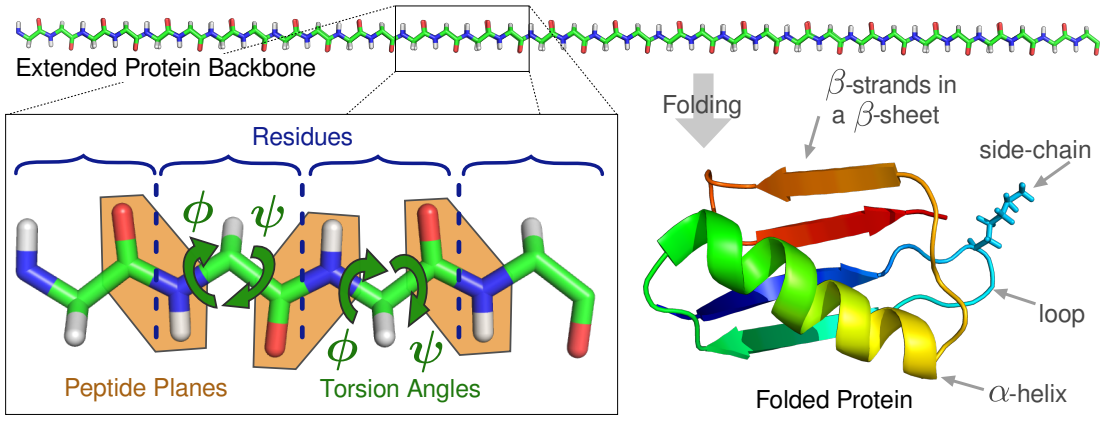


FIGURE 1.1: Structural features common to all proteins.

information about the mechanisms behind the observed functions. The goals of these structure-based studies are two-fold. If the mechanism of viral membrane fusion can be understood in sufficient detail, then 1) there is hope to design a molecule (i.e. a drug) that can interfere with membrane fusion and prevent infection by HIV-1 and 2) a suitable antigen can be constructed that both mimics HIV-1 envelope proteins and also elicits production of neutralizing antibodies against HIV-1 in the human immune system. The latter goal effectively seeks a vaccine which would equip the human immune system with the tools to actively fight an HIV-1 infection.

Often, an experimental challenge for structure determination is to devise a derivative of the protein of interest (i.e. a *construct*) suitable for experimental study. In its native and isolated form, the protein of interest might be nonreactive, unstable, insoluble, or otherwise uninteresting. Therefore, when investigating intermediate phases of a multi-step chain reaction (e.g. HIV-1 viral membrane fusion), the reactions must be arrested at the desired phase to preserve the protein of interest in its temporary conformation. Then, a protein construct must be designed that stabilizes the protein in this temporary conformation under experimental conditions so it may endure through the subsequent data collection experiments. Structure determination efforts then proceed on the protein construct under the assumption that observations

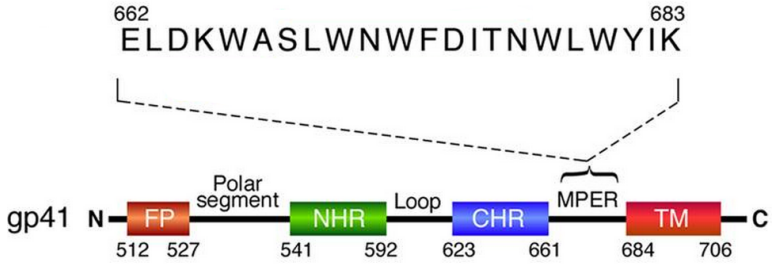


FIGURE 1.2: Amino acid sequence of MPER in the protein gp41. The fusion peptide (FP), N-terminal heptad repeat α-helix (NHR), C-terminal heptad repeat α-helix (CHR), and transmembrane (TM) regions of gp41 are also shown. Figure modified from Song et al. (2009).

from the construct are also applicable to the original protein.

gp41 is thought to play an important role in viral membrane fusion, but is a difficult target to study since it undergoes large conformational rearrangements and is also highly dependent on neighboring proteins gp121 (on the viral envelope) and CD4 receptors (on the host cell) to perform its function (Si et al., 2004). Hence, isolating the protein in an active form for structure determination is a difficult task. One such gp41 construct, designed by collaborators Dr. Patrick Reardon and Prof. Len Spicer at Duke, focused on the Membrane Proximal External Region of gp41 (MPER (Song et al., 2009), see Figure 1.2) which is hypothesized to function in tandem with two other identical copies of itself as a *trimeric* protein complex. To stabilize the MPER fragment of gp41 in the trimeric form, Foldon, a highly stable trimerizing protein domain from bacteriophage T4 fibritin (Tao et al., 1997), was genetically attached to MPER with a four-residue flexible linker and expressed as a fusion protein. Then, structure determination proceeded on the MPER-Foldon construct using Nuclear Magnetic Resonance spectroscopy (NMR).

1.2 Nuclear magnetic resonance spectroscopy

Of the two primary methods able to determine protein structures at atomic resolution, NMR is the only one that can interrogate structural information from proteins

in the physiologically-relevant solution state. X-ray crystallography, on the other hand, is often considered a gold standard for protein structure determination, but can produce structural artifacts that arise from packing the molecule into a crystal lattice. In contrast to X-ray crystallography methods, which are often viewed as frozen "snapshots" of a molecule, NMR is able to record evidence of the motions of proteins (*dynamics*) over time. Extracting protein dynamics from NMR experiments, while not the focus of this work, is the focus of a large body of work in the NMR community (Lange et al., 2008; Long and Brüschweiler, 2011; Tolman and Ruan, 2006; Tripathy et al., 2011) and generally requires a wealth of experimental NMR data. In this work, since we attempt to minimize the amount of NMR data needed for structure determination, we focus on the accurate reconstruction of static solution-state protein structures at atomic resolution.

One of the first steps in any structural study by NMR is the measurement and assignment of nuclear *resonances*. Each spin $\frac{1}{2}$ nucleus in the protein sample (which includes ^{1}H protons, and the less abundant isotopes ^{13}C and ^{15}N) resonates in a strong magnetic field at a particular frequency. Resonances manifest in NMR spectra as peaks in the frequency domain after Fourier analysis. When compared to a standard resonance frequency, each observed resonance frequency yields a *chemical shift* for the originating nucleus. Chemical shift values are also frequencies, but due to their tiny magnitude, they are often expressed as parts per million (ppm) or even parts per billion. In higher-dimensional NMR spectra, each peak is described by multiple chemical shifts. For example, in a 2D ^{15}N HSQC spectrum, each resonance peak is described by two chemical shifts: one in the ^{1}H dimension, and one in the ^{15}N dimension (See Figure 1.3a). Chemical shifts are also not necessarily unique to each nucleus since there can be many overlapping peaks and ambiguity in the NMR spectra. Hence, one of the first challenges to structure determination by NMR is finding the mapping between the experimentally-measured chemical shifts and the

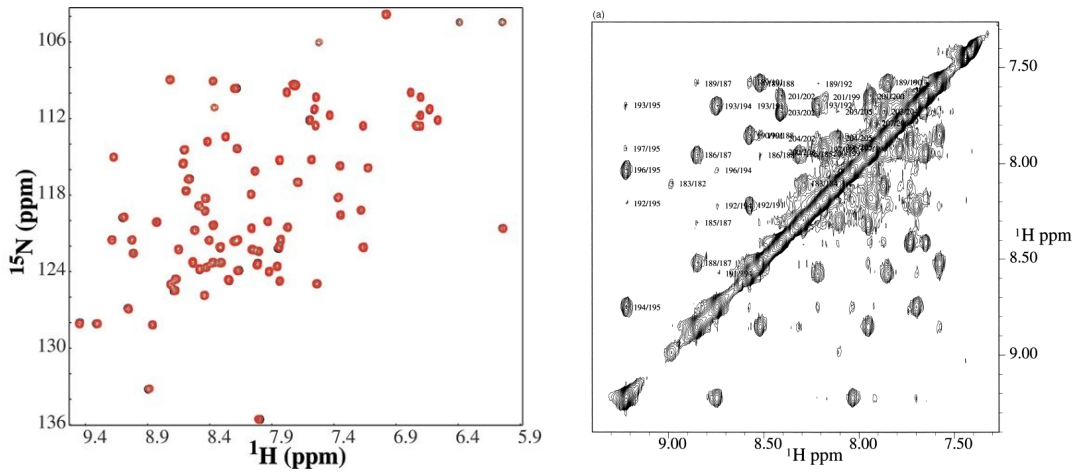


FIGURE 1.3: (a): Two nearly-identical, superimposed, and unlabeled ^{15}N/^{1}H-HSQC spectra of Human Ubiquitin showing peaks with two chemical shifts each: one in the ^{1}H dimension and one in the ^{15}N dimension. Figure reproduced from Ozkan et al. (2005). (b): The ^{1}H-^{1}H 2D NOESY spectrum of the natural abundance (i.e. unlabeled) L180-E207 fragment of the C protein showing diagonal peaks and cross peaks. Figure reproduced from Shahied et al. (2001).

atoms known from the amino acid sequence of the protein. This assignment can be typically performed for atoms along the protein backbone using 3D resonance experiments that tether together adjacent pairs of backbone resonances (Coggins and Zhou, 2003). The entire mapping of backbone resonances to the protein sequence is found by chaining the pairs into order by their overlapping chemical shifts.

Once chemical shifts are recorded and assigned, often the next step is to record NMR spectra exploiting the Nuclear Overhauser Effect (NOESY spectra). These spectra relate two proton resonances via a third peak (a *cross peak*) that shares the chemical shift values of the proton resonance peaks. The intensity of each NOESY cross peak is used to derive bounds on the distance between its referenced pair of protons. These distance restraints (*NOEs*) are used as direct geometric constraints on the protein structure. In theory, the NOESY cross peaks share the chemical shift values of the proton resonances, but in practice, experimental uncertainty renders

the assignment of cross peaks to proton resonances (*NOE assignment*) a difficult task. Proton resonances can overlap and as a result, a NOESY cross peak may appear to reference several proton resonances, hence giving rise to ambiguous NOE assignments.

NOE assignment ambiguity is a significant hurdle to structure determination efforts (Nilges and O'Donoghue, 1998; Rieping et al., 2007; Linge et al., 2004, 2001), and although experimental techniques such as 3D and 4D NOESY (Marion et al., 1989; Kay et al., 1990) (optimized with sparse time domain sampling (Coggins et al., 2010; Werner-Allen et al., 2010)) can alleviate many of the spectral degeneracies, current NOESY methodology is unable to completely resolve NOE assignment ambiguity for protein complexes with three or more identical subunits (Martin et al., 2011c,a; O'Donoghue et al., 2000). A NOESY spectrum might unambiguously show which proton within each subunit is referenced by a particular cross peak (i.e. completely resolving *atom ambiguity*), but it is not currently possible to determine in which subunit the referenced proton lies. This *subunit ambiguity* presents a formidable challenge for structure determination of protein *homo-oligomers*, or protein complexes with some fixed number of identical subunits. Isotopically-filtered NOESY (Ikura and Bax, 1992) can separate the intermolecular NOEs from the intramolecular NOEs, which is sufficient to determine the structure of a dimeric protein complex, but it cannot resolve subunit ambiguity for trimers or higher-order complexes. With no current experimental technique to resolve subunit ambiguity for trimers and higher-order complexes, there is a clear need for computational techniques to address this gap.

Another source of structural information derived from NMR experiments is the Residual Dipolar Coupling (RDC). The interpretation of RDC data is much more straightforward than the interpretation of NOESY spectra since there is no need to assign cross peaks. By comparing changes in NMR spectra between experimental

conditions where the protein tumbles freely (isotropically) in solution, and where the protein is weakly oriented (anisotropically) in the solution through interactions with an aligning media, a specified set of internuclear vectors in the protein (e.g. NH bonds) yield a set of scalar RDC values. An RDC value d, reports on the orientation of a specific bond vector in the protein through the relation:

$$d = d_{\max} \mathbf{v}^T S \mathbf{v} \tag{1.1}$$

where where $d_{\max}$ is the dipolar interaction constant (which subsumes various physical constants), $\mathbf{v}$ is the internuclear vector orientation relative to an arbitrary molecular frame, and S is the *alignment tensor*, a 3×3 real matrix that is both symmetric and traceless (Donald, 2011). The alignment tensor represents the average alignment of the protein in the anisotropic environment. Once decomposed into its constituent rotation and scaling components, the alignment tensor and the RDC values can be used to solve analytically for the internuclear vectors. RDCs, along with NOEs and chemical shifts, are the primary sources of information from NMR experiments that are used to derive geometrical constraints for atomic resolution protein structure determination.

2

DISCO: A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs

The text of this chapter has been adapted from published manuscripts that were co-authored with A. K. Yan, C. Bailey-Kellogg, P. Zhou, and B. R. Donald. In this section, my primary contribution was implementing DISCO, applying DISCO, and collaborating on the design of the DISCO algorithm.

J. W. Martin, A. K. Yan, C. Bailey-Kellogg, P. Zhou, and B. R. Donald. A graphical method for analyzing distance restraints using residual dipolar couplings for structure determination of symmetric protein homo-oligomers. Protein Science, 2011. 20(6):970–985.

J. W. Martin, A. K. Yan, C. Bailey-Kellogg, P. Zhou, and B. R. Donald. A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs. Journal of Computational Biology, 2011. 18(11):1507–1523.

Abstract: High-resolution structure determination of homo-oligomeric protein complexes remains a daunting task for NMR spectroscopists. Although isotope-filtered experiments allow separation of intermolecular NOEs from intramolecular NOEs and determination of the structure of each subunit within the oligomeric state, degenerate chemical shifts of equivalent nuclei from different subunits make it difficult to assign intermolecular NOEs to nuclei from specific pairs of subunits with certainty, hindering structural analysis of the oligomeric state. Here, we introduce a graphical method, DISCO, for analysis of intermolecular distance restraints and structure determination of symmetric homo-oligomers using residual dipolar couplings. Based on knowledge that the symmetry axis of an oligomeric complex must be parallel to an eigenvector of the alignment tensor of residual dipolar couplings, we can represent distance restraints as annuli in a plane encoding the parameters of the symmetry axis. Oligomeric protein structures with the best restraint satisfaction correspond to regions of this plane with the greatest number of overlapping annuli. This graphical analysis yields a technique to characterize the complete set of oligomeric structures satisfying the distance restraints, and to quantitatively evaluate the contribution of each distance restraint. We demonstrate our method for the trimeric *E. coli* Diacylglycerol Kinase, addressing the challenges in obtaining subunit assignments for distance restraints. We also demonstrate our method on a dimeric mutant of the immunoglobulin-binding domain B1 of streptococcal protein G to show the resilience of our method to ambiguous atom assignments. In both studies, DISCO computed oligomer structures with high accuracy despite using ambiguously-assigned distance restraints.

2.1 Introduction

A vast number of macromolecules, including many membrane proteins in higher eukaryotic cells, form symmetrical oligomeric complexes containing multiple subunits

(Goodsell and Olson, 2000). The determination of high-resolution solution structures of oligomeric protein complexes, unfortunately, remains a difficult task (White, 2004). In NMR studies, the fundamental challenge for such systems is that the equivalent atoms from different subunits share identical chemical shifts. Therefore, even if it is possible to narrow down the observed NOEs to particular pairs of nuclei, it is still difficult to determine within which subunits these nuclei are located. This dilemma can be partially resolved by isotope-filtered experiments and elegant isotope labeling schemes, which have made it possible to isolate intermolecular NOEs. By excluding intermolecular NOEs from the complete set of NOE distance restraints, it is possible to determine the high-resolution structure of each subunit based on entirely intramolecular restraints (Oxenoid and Chou, 2005; Schnell and Chou, 2008; Wang et al., 2009). However, current techniques are not able to differentiate which pairs of subunits contribute to the observed intermolecular dipolar interactions, giving rise to *subunit ambiguity* (Potluri et al., 2007) (See Figure 2.1). Subunit ambiguity hinders analysis of not only NOEs, but also distance restraints derived from disulfide bonds. While identical chemical shifts for symmetric protons hinders subunit assignment, merely similar chemical shifts also complicate resonance assignment. *Atom ambiguity* (Potluri et al., 2007) characterizes an NOE that could be assigned to multiple pairs of nuclei when overlapping ranges of chemical shifts are unable to be resolved. Therefore, NOE assignment and structure determination of trimeric complexes or complexes with higher-order symmetry remains an unresolved challenge for NMR spectroscopists.

Even with precise unambiguous distance restraint assignments, structure determination remains a difficult task. Structure determination protocols that rely on distance geometry calculations are computationally-expensive to perform on large proteins, and protocols that rely on simulated annealing require careful selection of annealing parameters, may not converge, or can potentially miss structures consistent

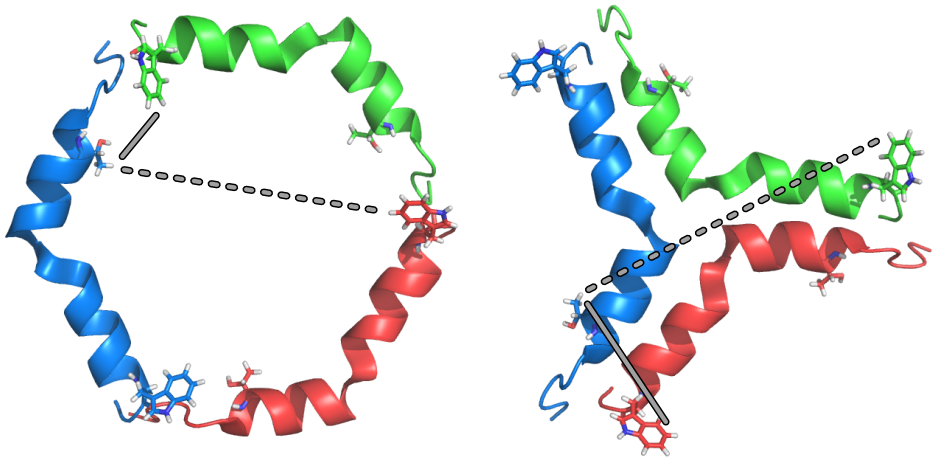


FIGURE 2.1: Subunit Ambiguity: An ambiguous intermolecular NOE between an H^{γ} proton of a Threonine residue and an $H^{\zeta 3}$ proton of a Tryptophan residue for a hypothetical symmetric trimer has two possible assignments. If we choose the H^{γ} proton to lie in the blue subunit, then the $H^{\zeta 3}$ proton could lie in either the green or red subunits. Therefore, the distance restraint either relates the blue-green pair of protons or the blue-red pair of protons. The choice of assignment can potentially lead to vastly different overall folds for the trimer. Left: A ring-shaped scaffold satisfies the blue-green assignment, but not the blue-red assignment. Right: A star-shaped scaffold satisfies the blue-red assignment, but not the blue-green assignment. Satisfied restraints are shown with solid grey lines. Unsatisfied restraints are shown with dashed grey lines.

with experimental restraints. However, by considering distance restraint assignment and oligomeric structure determination simultaneously, we arrive at an elegant solution. Our approach addresses both assignment and structure determination by incorporating information from residual dipolar couplings (RDCs) into the analysis.

In traditional and widely-used NMR structure determination protocols for symmetric homo-oligomers, RDCs are typically saved until the structure refinement phase, after calculation of an initial fold using a combination of distance restraints and restraints on dihedral angles. Work by Nilges (Nilges, 1993) focused on calculating oligomer models where additional potentials guided the protein structure to satisfy the symmetry constraints. The method relied primarily on simulated annealing and molecular dynamics, and has been successfully employed in structure determination of homo-oligomers including a trimer (Kovacs et al., 2002) and a hex-

amer (O'Donoghue et al., 2000). A non-crystallographic symmetry potential ensured subunits shared the same local conformation modulo relative placement and global orientation, while an additional potential arranged the subunits symmetrically by minimizing differences in distances for a chosen subset of the distance restraints. When it is difficult to assign distance restraints unambiguously, ARIA (Rieping et al., 2007) can be used to perform simultaneous structure calculation and distance restraint assignment using ambiguous distance restraints, and has been improved by Bardiaux et al. (Bardiaux et al., 2009) who implemented network anchoring (Herrmann et al., 2002) and spin-diffusion correction (Linge et al., 2004) into the framework.

We propose a new protocol for structure calculation of homo-oligomers with cyclic symmetry that incorporates RDCs into the beginning of the oligomeric assembly method, so that we may take advantage of the global nature of the restraint provided by the RDCs. Our RDC-first approach creates a framework in which we analyze local intermolecular distance restraints without requiring a complete oligomer structure. Instead, the oligomer structure can be represented in terms of its axis of symmetry and the structure of its subunit (see Figure 2.2). Therefore, we perform structure determination in the configuration space of symmetry axes: two translational degrees of freedom (a plane, $\mathbb{R}^2$) and two rotational degrees of freedom (a unit sphere, $\mathbb{S}^2$).

Our method, DISCO, uses the observation that the symmetry axis of the oligomeric structure must be parallel to one of the eigenvectors of the alignment tensor (Al-Hashimi et al., 2000), and therefore uses RDCs to compute the orientation (in $\mathbb{S}^2$) of the symmetry axis. Uncertainty in the orientation of the symmetry axis (due to experimental error) is considered by perturbing the experimental RDC values via sampling from a normal distribution, a basic technique that has been previously used to model the experimental error of RDCs for backbone structure determination of monomers (Donald and Martin, 2009; Wang and Donald, 2004; Zeng et al.,

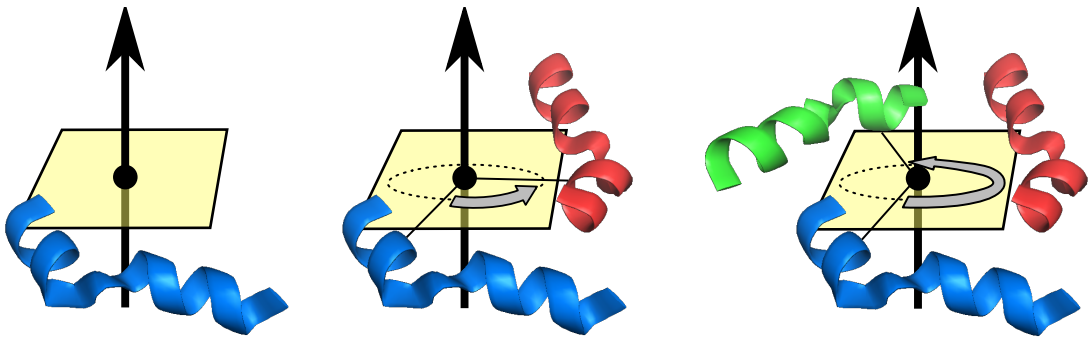


FIGURE 2.2: Generating a Trimer Structure using Symmetry: Left: Compute the position and orientation of the symmetry axis (vertical arrow) relative to the subunit structure (blue α-helix). Middle: Copy the subunit structure and rotate by 120° about the symmetry axis to place the second subunit (red α-helix). Right: Copy the subunit structure again and rotate by 240° to place the final subunit (green α-helix).

2009). DISCO computes a set of sampled RDC values by drawing a single sample from a normal distribution for each recorded RDC value. Using the set of sampled RDC values and the subunit structure, DISCO computes an alignment tensor which describes a possible orientation for the symmetry axis. By computing a large number of alignment tensors from perturbed RDCs, it is possible to estimate the set of possible symmetry axis orientations, which DISCO conservatively bounds, and then approximates using a systematic grid search (See Methods, Section 2.3.2).

For each orientation on the grid (a *grid orientation*), DISCO uses experimental intermolecular distance restraints such as NOEs and disulfide bonds to compute the position of the symmetry axis – even when precise subunit assignments and atom assignments are not known. Using the computed orientation of the symmetry axis, each possible assignment for a distance restraint restricts the positions of the symmetry axis to an annulus in the plane ($\mathbb{R}^2$). Each annulus is the set of points lying between two concentric circles whose two radii are mathematically derived from the upper and lower bounds of the corresponding distance restraint (See Section 2.3.4 and Figure 2.12). Both the inner and the outer radii of each annulus are also dependent on uncertainty in the subunit structure (See Section 2.3.3). Multiple possible

assignments for a distance restraint are encoded as a union of the annuli resulting from each possible assignment. DISCO analyzes the unions of annuli using a geometric algorithm (Section 2.3.5) to compute the *maximally satisfying regions* (*MSRs*) of the plane, each of which defines a continuous set of symmetry axis positions representing the complete set of oligomer structures that satisfy the greatest number of intermolecular distance restraints.

Specifically, let us refer to any oligomer structure whose symmetry axis orientation has been computed from RDCs as an *oriented oligomer structure*. Therefore, the space of oriented oligomer structures corresponds to the space of symmetry axis positions. In the case that all distance restraints are simultaneously satisfiable, DISCO can guarantee the MSRs describe all satisfying oriented oligomer structures without missing any of them. If all distance restraints cannot be satisfied, DISCO can guarantee that any oriented oligomer structure whose symmetry axis position has been chosen from the MSRs will satisfy strictly more intermolecular distance restraints than oriented oligomer structures whose symmetry axis positions have been chosen from outside the MSRs.

Previous work (Wang et al., 2008; Potluri et al., 2006, 2007) also formulated structure determination of homo-oligomers in a symmetry configuration space. Potluri et al. (Potluri et al., 2006, 2007) computed the orientation and position of the symmetry axis without RDCs using hierarchical subdivision of the configuration space ($\mathbb{R}^2 \times \mathbb{S}^2$). The configuration space was partitioned into regions which were pruned if geometric bounds proved they did not contain any symmetry axes whose oligomer structures satisfied the intermolecular NOEs. Otherwise, the regions were subdivided and the search recursed on their children. Wang et al. (Wang et al., 1998) computed symmetry parameters for oligomer models using ambiguously-assigned distance restraints by partitioning Cartesian space instead of axis configuration space. After choosing three of the distance restraints as a geometric base, AMBIPACK (Wang

et al., 1998) computed symmetry axis parameters by computing the rigid transformation across the interface between two identical subunits. The three chosen distance restraints were used to define a coarse relative orientation between the subunits at the interface, which was iteratively refined against the remaining distance restraints. However, since random sampling and numerical optimization were used to calculate geometric bounds, AmbiPack is unable to guarantee that all satisfying structures will be discovered. Wang et al. (Wang et al., 2008) computed the orientation of the symmetry axis using just RDCs. The axis position was computed by generating putative dimer models on a grid over $\mathbb{R}^2$ and scoring the intermolecular interface using a residue-pairing molecular mechanics function. Since dimer models were ranked only according to molecular mechanics scores, van der Waals energy, and agreement with the RDCs, the method does not incorporate the structural information provided by intermolecular distance restraints into the calculation.

disco computes symmetry parameters explicitly by analyzing RDCs and distance restraints such as NOEs and disulfide bonds. disco computes dimer models and also generalizes to trimers and higher-order oligomers by considering subunit ambiguity. Possible subunit and atom assignments for an intermolecular distance restraint are encoded as a union of annuli in $\mathbb{R}^2$, allowing our method to analyze all assignments simultaneously and avoid the need for explicit (and expensive) enumeration of possible assignment combinations. Furthermore, all distance restraints are given the same geometric treatment, avoiding the need to subjectively select a small number of distance restraints at the outset to bootstrap the structure determination. Representing distance restraints as planar annuli also allows us to analyze each restraint independently. We characterize a distance restraint as *inconsistent* if its corresponding union of annuli does not contain any of the MSRs. No oriented oligomer structure whose symmetry axis position was chosen from a MSR could satisfy an inconsistent restraint. Moreover, disco can compute the MSRs exactly

without relying on random sampling or numerical optimization, and therefore is able to guarantee that no satisfying oriented oligomer structures will be missed.

To demonstrate the ability of DISCO to perform structure determination without subunit assignments, we show results for *E. coli* Diacylglycerol Kinase (DAGK) (Van Horn et al., 2009) using disulfide bonds as distance restraints in Section 2.2.1. Like intermolecular NOEs, subunit assignments for disulfide bonds are not known. In addition to considering subunit ambiguity, DISCO also considers atom ambiguity. To demonstrate the resilience of DISCO under ambiguous atom assignments for NOEs, we show results for a dimeric mutant of the immunoglobulin-binding domain B1 of streptococcal protein G (the GB1 domain-swapped dimer) (Byeon et al., 2003) in Section 2.2.2. The GB1 mutant differs from the wild type by the L5V/F30V/Y33F/A34F mutations resulting in a domain-swapped dimer. Section 2.3 describes the methodology for our computational tests.

2.2 Results and discussion

2.2.1 Structure determination of DAGK under subunit ambiguity

To compute the oligomeric structure for the trimeric DAGK, we used the following experimental data: 67 NH RDCs and 24 disulfide bonds per subunit (Van Horn et al., 2009). We chose model 1 from PDB (Berman et al., 2000) ID: 2KDC (Van Horn et al., 2009) to serve as the reference structure. The subunit structure used by DISCO was the first subunit in the reference structure, which was determined using traditional protocols. This mirrors the situation where the subunit structure can be determined with confidence (Oxenoid and Chou, 2005; Schnell and Chou, 2008; Wang et al., 2009), but the main bottleneck is subunit assignment and the assembly of subunit structures to form the oligomer structure.

To compute an initial coarse orientation for the symmetry axis, DISCO first computes an alignment tensor from the RDCs and the subunit structure (See Section 2.3.1

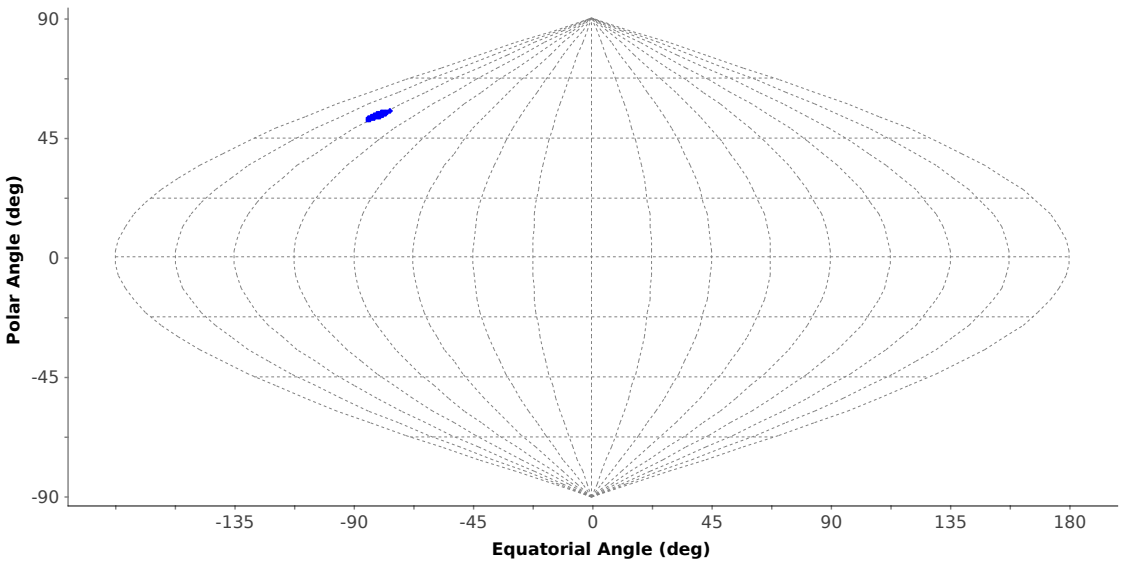


FIGURE 2.3: Sampling 10,000 sets of perturbed values using the experimental NH RDCs for DAGK resulted in the symmetry axis orientations (blue) shown on a sinusoidal or Sanson-Flamsteed projection (Bugayevskiy and Snyder, 1995).

for more details). The rhombicity of the computed alignment tensor is 0.02, which is near zero, the value one expects for a symmetric trimer. The alignment tensor fits well to the RDCs and the subunit structure, which is shown by computing the RMS deviation of the recorded RDC values to those back-computed from the subunit structure (the *RDC RMSD*, 0.28 Hz for the NH RDCs). Using 10,000 sets of sampled RDCs (See Section 2.3.2), DISCO computed 10,000 alignment tensors whose D_{zz} eigenvectors (the z-axes, using the notation of Clore et al. (Clore et al., 1998) and Wedemeyer et al. (Wedemeyer et al., 2002)) show possible symmetry axis orientations (shown in Figure 2.3). The RDC values were sampled from normal distributions with standard deviations equal to 1 Hz, resulting in sampled RDC values differing from the recorded RDC values by as much as 4.7 Hz, which are significant deviations for NH RDCs. Figure 2.4 shows the symmetry axis orientation for the reference structure which is within the range of the z-axes resulting from the RDC sampling, and also shows the grid of orientations used by DISCO to approximate the z-axes.

DISCO computed MSRs for each of the 17 grid orientations (from Figure 2.4),

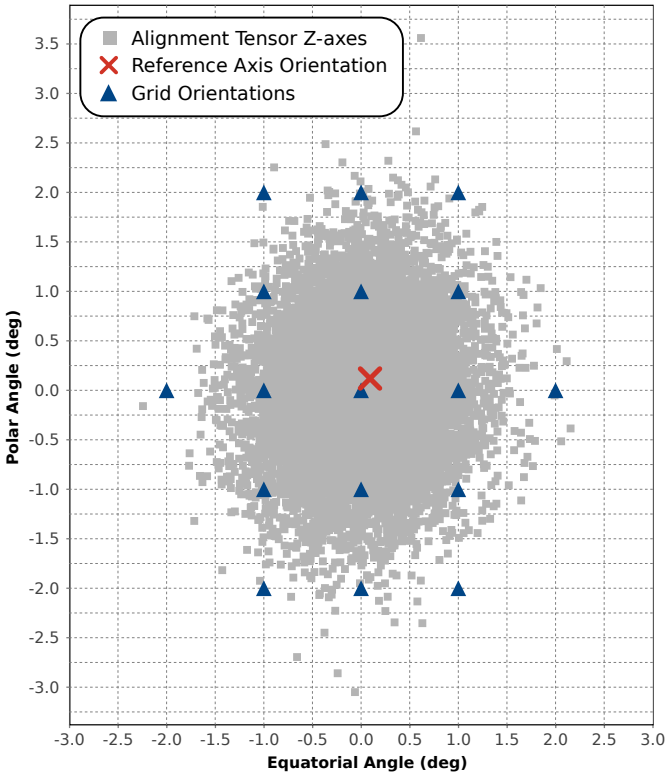


FIGURE 2.4: A comparison of the reference symmetry axis orientation (red) against the 10,000 z-axes resulting from RDC sampling (grey) for DAGK. The orientations illustrated by the blue triangles (the grid orientations) were each used, in turn, to generate constraint annuli.

drawn from a uniform grid with a resolution of 1°. Figure 2.5 shows the MSR computed from the disulfide bonds for the most central grid orientation (at coordinates (0,0) in Figure 2.4). The MSRs contain the position of the symmetry axis for the reference structure, indicating that the distance restraint analysis is able to successfully recover the symmetry parameters of the reference structure. Since DISCO computes the exact set of oriented oligomer structures consistent with the distance restraints, the absence of any additional MSRs farther away rules out the possibility of a satisfying oligomer structure that is dissimilar to those already discovered by the algorithm.

In order to perform detailed structural analysis, we generate a set of discrete structures to represent the MSRs by sampling symmetry axis positions from the

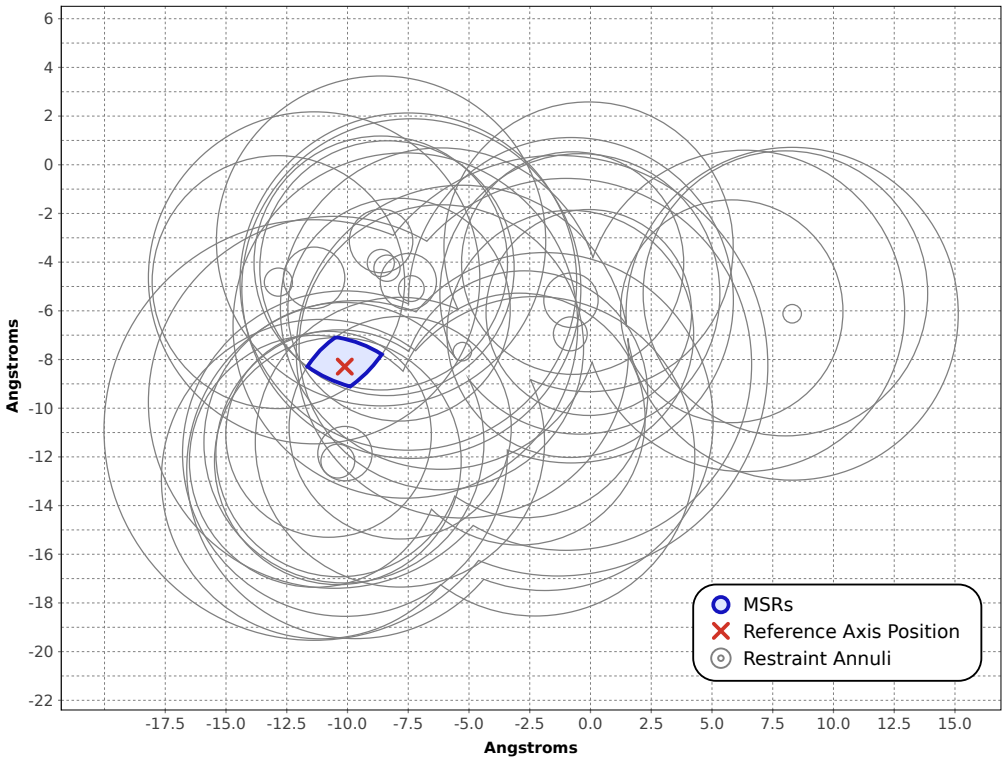


FIGURE 2.5: Distance restraint unions of annuli and MSR computed from the 24 disulfide bonds for DAGK using the central grid orientation from Figure 2.4. This MSR is one of 17 computed for DAGK.

MSR interiors. One of the advantages of DISCO is that by computing the exact MSRs, it is unnecessary to sample the entire symmetry axis position configuration space. Instead, we can sample only within the MSRs at a much finer resolution than would be possible using a grid search over the full configuration space. This is especially important when separate MSRs are computed for each symmetry axis orientation represented by the grid points. Symmetry axis positions were sampled from the MSRs from the 17 grid orientations on a 0.75 Å resolution uniform grid resulting in 68 oligomer structures for DAGK. All oligomer structures computed by DISCO were within 1.5 Å backbone atom RMSD to the reference structure, with the closest at 0.3 Å.

We scored computed structures according to two criteria (Potluri et al., 2007):

RMS distance restraint violation and van der Waals energy (measured in kcal/mol) after energy-minimization in X-PLOR (Schwieters et al., 2003) where the backbone remains fixed, but side chains are allowed to re-pack. The RMS distance restraint violation measure scores structural agreement with the intermolecular distance restraints, and the van der Waals energy scores the structures for intermolecular packing. Since DISCO can discriminate between consistent and inconsistent distance restraints, it is easily possible to minimize the computed oligomer structures subject to only the consistent distance restraints, ensuring that inconsistent restraints cannot influence the final minimized structures. However, all of the disulfide bonds were mutually consistent and hence, the minimization was conducted with all available distance restraints. Figure 2.6 plots the scores of all computed structures for DAGK as well as the score for the reference structure. The computed structures have distance restraint satisfaction scores distributed around the score of the reference structure, with computed structures scoring as much as 0.12 Å better. Since we expect an oligomer structure to have better packing than the subunit alone, the six structures with energies higher than the van der Waals energy of the subunit in isolation (-367 kcal/mol) were removed from the final computed ensemble. Figure 2.7 shows all 68 oligomer structures computed by DISCO aligned to the reference structure.

Since DISCO can compute the complete set of oriented oligomer structures consistent with the distance restraints, the average RMS deviation from the mean for each backbone atom of the computed structural ensemble represents uncertainty about the position of the symmetry axis inherent in the distance restraints. Structure determination methods that can fail to report satisfying oligomeric conformations (possibly due to under-sampling) can only report the RMS deviation from the mean for each atom of the computed ensemble of structures, which is unable to completely characterize uncertainty about the symmetry axis position. DISCO computed an average RMS deviation from the mean of 1.12 Å for all atoms and 1.08 Å for backbone

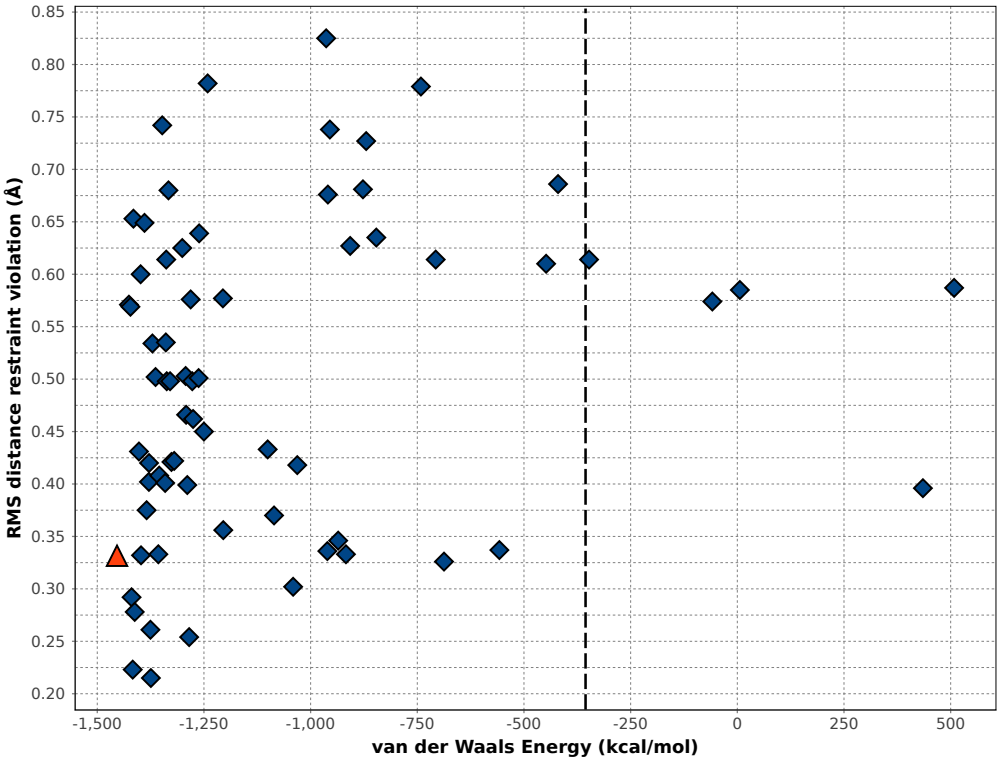


FIGURE 2.6: Distance restraint satisfaction scores (lower is better) and van der Waals energies for computed (blue) and reference (red) oligomer structures for DAGK after minimization. The energy cutoff at -367 kcal/mol is shown with a black dashed line. One computed structure with a very high van der Waals energy has been omitted from the figure.

atoms for the 68 minimized oligomer structures for DAGK.

2.2.2 Structure determination of the GB1 domain-swapped dimer under atom ambiguity

To compute the dimeric structure of the GB1 domain-swapped dimer, we used 56 NH RDCs and 296 experimental intermolecular NOEs (initially assigned unambiguously) per subunit (Byeon et al., 2003). We chose model 1 from PDB ID: 1Q10 (Byeon et al., 2003) to serve as the reference structure. The subunit structure used by DISCO was the first subunit in the reference structure, which was determined using traditional protocols. Again, DISCO focuses on the oligomeric assembly bottleneck since the subunit structures can in many cases be determined with confidence (Oxenoid and

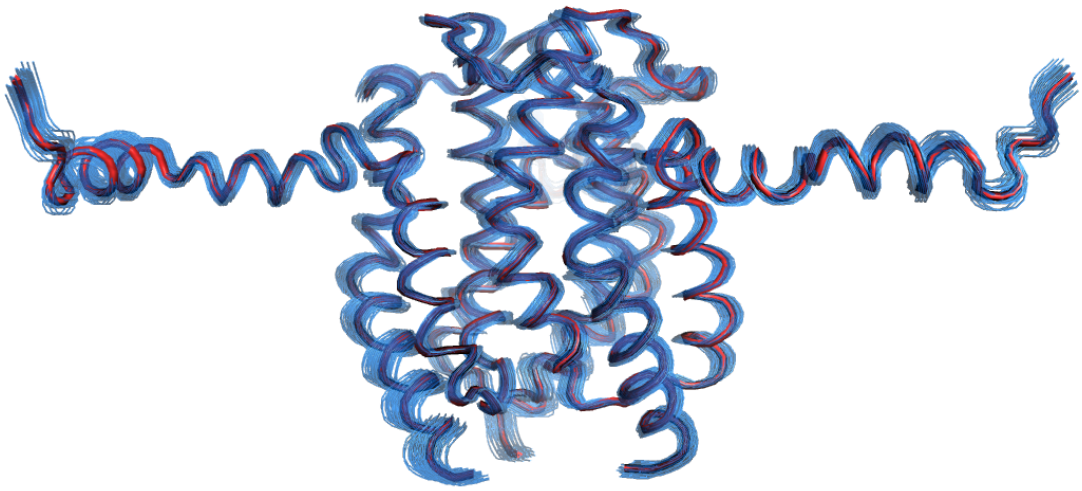


FIGURE 2.7: The 68 oligomer structures computed by DISCO (thin blue strands, including the 6 removed by the energy cutoff) are all within 1.5 Å backbone atom RMSD to the reference (larger red backbone) for DAGK.

Chou, 2005; Schnell and Chou, 2008; Wang et al., 2009). While the symmetry axis for a dimer must be parallel to one of the eigenvectors of the alignment tensor, *which* eigenvector satisfies this condition cannot be uniquely determined from RDCs alone. A search over the three possible choices revealed the D_{xx} eigenvector as the best candidate (see Section 2.3.1 for more details about this search). The alignment tensor computed from the recorded RDCs and the subunit structure fits well, with a RDC RMSD of 0.57 Hz.

Using 10,000 sets of sampled RDCs (See Section 2.3.2), DISCO computed 10,000 alignment tensors whose x-axes show possible symmetry axis orientations. The RDC values were sampled from normal distributions with standard deviations equal to 1 Hz, resulting in sampled RDC values differing from the recorded RDC values by as much as 4.6 Hz. Similarly to DAGK, the symmetry axis orientation for the reference structure for the GB1 domain-swapped dimer was also within the range of x-axes resulting from the RDC sampling. DISCO computed MSRs for each of the 19 grid orientations, which were drawn from a 0.75° resolution uniform grid. Figure 2.8 shows

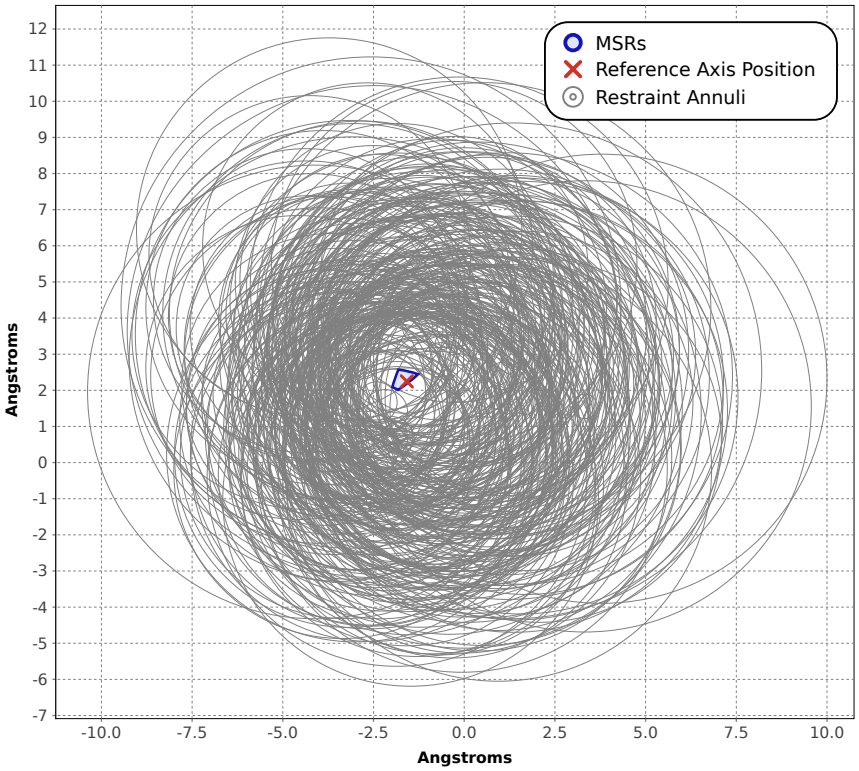


FIGURE 2.8: Distance restraint unions of annuli and MSR computed from the 296 NOEs for the GB1 domain-swapped dimer using the central grid orientation. This MSR is one of 19 computed for the GB1 domain-swapped dimer.

the single MSR computed from the NOEs for the most central grid orientation.

DISCO sampled the 19 MSRs from the grid orientations at a resolution of 0.25 Å which produced 48 oligomer structures, all of which were within 0.72 Å backbone RMSD to the reference with the closest at 0.07 Å. Figure 2.9 shows the scores for the energy-minimized oligomer structures compared to a minimized version of the reference structure. The backbone was fixed during minimization, but sidechains were allowed to re-pack and all available NOEs were used to restrain the oligomer structures, since DISCO did not discover any inconsistent NOEs. The structures computed for the GB1 domain-swapped dimer have distance restraint satisfaction scores distributed around the score for the reference structure, with some computed structures scoring negligibly (almost 0.03 Å) better. A single computed structure scored

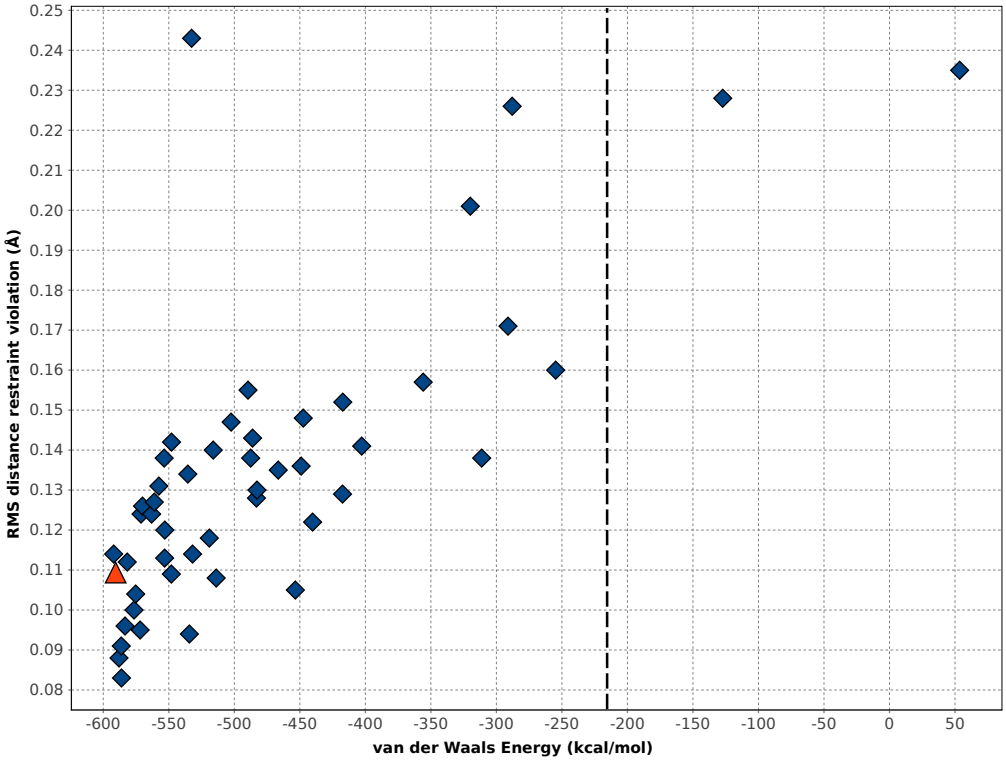


FIGURE 2.9: Distance restraint satisfaction scores (lower is better) and van der Waals energies for computed (red) and reference (blue) oligomer structures for the GB1 domain-swapped dimer.

with a lower van der Waals energy than the reference structure (by a narrow margin of 1.5 kcal/mol), and most of the remaining computed structures scored within 200 kcal/mol of the reference. The energy cutoff (-215 kcal/mol) was determined as in Section 2.2.1. Two structures whose van der Waals energies were over the energy cutoff were removed from the final computed ensemble. Figure 2.10 shows all 48 of the oligomer structures computed by DISCO aligned to the reference structure. The DISCO ensemble has an average RMS deviation from the mean of 0.54 Å for all atoms and 0.50 Å for backbone atoms for the 48 computed oligomer structures for the GB1 domain-swapped dimer.

DISCO analyzes distance restraints with atom ambiguity as well as subunit ambiguity. All of the NOEs for the GB1 domain-swapped dimer were deposited as

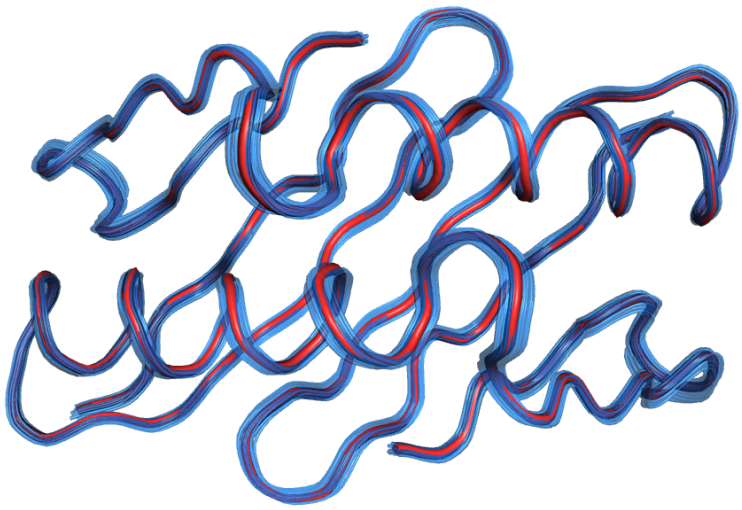


FIGURE 2.10: The 48 oligomer structures computed by DISCO (thin blue strands, including the 2 removed by the energy cutoff) are all within 0.72 Å backbone atom RMSD to the reference (larger red backbone) for the GB1 domain-swapped dimer.

unambiguously-assigned restraints (Byeon et al., 2003), so we simulated atom ambiguity by expanding the assignments to include protons with similar chemical shifts. We obtained 294 ^{1}H, 175 ^{13}C, and 61 ^{15}N chemical shifts from the BMRB (Ulrich et al., 2007) using the accession number 5875. We chose a window size of 0.05 ppm for Hydrogen shifts (δ_H), and a window size of 0.5 ppm for the Nitrogen and Carbon shifts (δ_Θ, See Section 2.3.8 for more details). This simulation increased the average number of assignments per NOE from 1 to 6.7.

To evaluate the effect of these additional NOE assignments on the range of oligomer structures computed by DISCO, we computed MSRs for only the central symmetry axis orientation, which was computed from the original recorded RDC values without perturbation. After comparison with the MSR computed from the same symmetry axis orientation, but using the original unambiguously-assigned NOEs (See Figure 2.11), we discovered the MSR computed from the expanded NOE assignments completely contains the MSR computed from the original NOE assignments, as well as the symmetry axis position of the reference structure. Structures sampled finely

from the single MSR computed from the expanded NOE assignments on a 0.05 Å resolution grid are all within 0.81 Å backbone RMSD to the reference. Since the MSR computed from the expanded NOE assignments is larger than the MSR computed from the original assignments, but still contains the reference axis position, these results indicate that despite a high degree of ambiguity in the distance restraints, DISCO still computes the correct symmetry axis positions – just at a slightly lower precision.

Of the 1909 expanded possible assignments for all the NOEs, DISCO discovered that 13.6% of them could not be satisfied by any oriented oligomer structure, indicating a conflict between the expanded possible assignments and the RDC-determined symmetry axis orientation. The annuli for this 13.6% of the expanded assignments enclosed no points (i.e., are the empty set) and therefore, no satisfying symmetry axis positions exist. Section 2.3.4 describes in more detail the distance restraint geometry that results in no satisfying symmetry axis positions for an assignment. These expanded assignments with no satisfying symmetry axes were clearly incorrect and were eliminated immediately using DISCO's RDC-first analysis.

2.3 Materials and methods

To perform the structure determination of DAGK and the GB1 domain-swapped dimer, we conducted a number of computational tests. The NMR data were downloaded from the PDB (Berman et al., 2000) and the BMRB (Ulrich et al., 2007); deposition IDs are given in Section 2.2; the data collection is described previously (Van Horn et al., 2009; Byeon et al., 2003). All computations were performed on a single core of an Intel Core i7 processor at 1.6 GHz which completed in time on the order of hours. The number and type of distance restraints are described in Section 2.2. To compute oligomer models, DISCO executes a seven-step protocol which is outlined in sections 2.3.1 to 2.3.7.

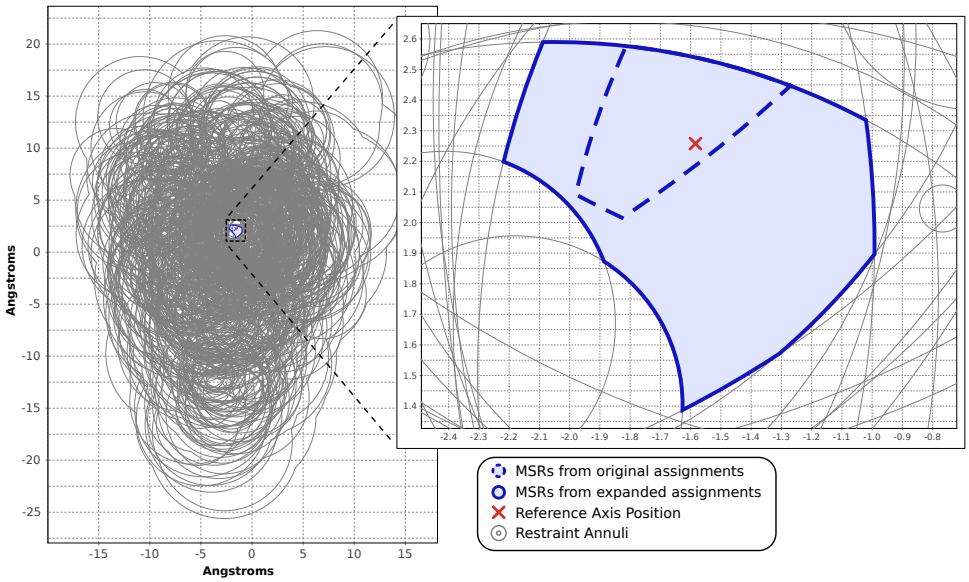


FIGURE 2.11: MSRs computed from the NOE assignments for the GB1 domain-swapped dimerwith simulated atom ambiguity. Structures sampled finely from the MSR (on a 0.05 Å resolution grid) are all within 0.81 Å backbone RMSD to the reference.

2.3.1 Computing the central symmetry axis orientation

DISCO considers uncertainty in the orientation of the symmetry axis by first computing a central orientation, and later perturbing it indirectly. The central orientation is computed from an alignment tensor fit to the recorded RDC values and the subunit structure using singular value decomposition (Losonczi et al., 1999). To evaluate the fit of this alignment tensor, DISCO back-computes RDCs for the subunit structure and computes the RMSD from the experimental RDCs. For homo-oligomers with cyclic symmetry, if the alignment tensor has zero rhombicity, one of its eigenvectors must be parallel to the symmetry axis (Al-Hashimi et al., 2000). Further details of computing the central orientation from an alignment tensor depends on the oligomeric state of the protein:

Trimers and higher-order oligomers: For trimers and higher-order oligomers, we expect an alignment tensor with zero rhombicity. In this case, the central symmetry axis is parallel to the eigenvector of the alignment tensor whose eigenvalue has the

largest magnitude (the principal axis, or z-axis). Alignment tensors with non-zero rhombicity for trimers and higher-order oligomers do not reflect the symmetry of the oligomer and we are unable to apply DISCO to the RDCs in that case.

Dimers: For dimers, any value of rhombicity is acceptable, although non-zero rhombicity is actually preferred, since it guarantees the alignment tensor will have at most three eigenvectors. Which eigenvector corresponds to the symmetry axis cannot be uniquely determined from a single set of RDCs alone, so all possibilities must be examined. If an alignment tensor has three distinct eigenvalues, as the tensor for the GB1 domain-swapped dimer does, then one needs to merely consider the three corresponding eigenvectors. DISCO evaluates each choice of eigenvector for distance restraint satisfaction by computing MSRs (see Section 2.3.5). The eigenvector whose MSRs satisfy the greatest number of distance restraints is selected as the central symmetry axis orientation.

2.3.2 Computing the uncertainty in the symmetry axis orientation

Once the central symmetry axis orientation has been computed, it is perturbed using the following method that uses the experimental error of the RDCs. For each recorded NH RDC value, define a normal distribution with mean equal to the RDC value, and standard deviation equal to 1 Hz. Experimental error for RDCs corresponding to different internuclear vectors can be modeled by varying the choice of standard deviation. Next, compute one set of sampled RDCs by sampling one value from each distribution. Then, fit an alignment tensor using the sampled set of RDCs and the subunit structure to compute one possible symmetry axis orientation. Repeat a large number of times (10,000 sufficed for our computational tests) to compute a large number of possible orientations. Next, bound the set of possible orientations within an elliptical cone. Finally, sample orientations uniformly from the elliptical cone at a desired resolution to compute the set of grid orientations.

DISCO uses the grid orientations to represent uncertainty in the orientation of the symmetry axis, and evaluates each grid orientation for agreement with the distance restraints in Sections 2.3.4 and 2.3.5.

2.3.3 Calculating subunit structural uncertainty

To account for uncertainty in the subunit structure, DISCO adds a padding value α_i to the upper and lower bounds of each distance restraint similarly to how NOEs are adjusted for pseudoatoms. The upper bound for a distance restraint D_i is increased by α_i and the lower bound is decreased by α_i. In the case the subunit structure was determined by NMR, the ensemble of structures for the subunit directly represents the uncertainty of each atom position. Alternatively (and for x-ray structures), simulations of molecular dynamics can be used to probe for structural variability in the subunit. DISCO computes a padding value for each atom involved in each possible assignment for the distance restraint. For a unique atom a in the subunit, let $E(a)$ be the set of all instances of that atom in the ensemble. Additionally, let $M(E(a))$, be a function that returns the maximum distance of any atom in $E(a)$ to the centroid of $E(a)$. Given a distance restraint $D_i = \{(\mathbf{p}_k, \mathbf{q}_k)\}$ relating two atoms $\mathbf{p}_k$ and $\mathbf{q}_k$ for each assignment k, DISCO computes $\alpha_i = \max_k M(E(\mathbf{p}_k)) + \max_k M(E(\mathbf{q}_k))$. Hence, the upper bound of the distance restraint increases with the uncertainty in the positions of the two related atoms, and the lower bound decreases towards zero. The computation of padding is automated and requires no user-defined parameters or human choices.

2.3.4 Computing distance restraint unions of annuli

For each possible assignment of each intermolecular distance restraint, DISCO computes one annulus which describes a continuous set of points in the configuration space of symmetry axis positions (a plane, $\mathbb{R}^2$). Each point in this annulus describes

an oriented oligomer structure that satisfies the assignment.

To compute an annulus for a distance restraint assignment, we first define a coordinate system in which the z-axis ($\hat{\mathbf{z}}$) is parallel to a chosen grid orientation. Using the structure of a single subunit A, we define the origin of this coordinate system to be the centroid of all the atoms in the subunit. DISCO computes the position of the symmetry axis in this coordinate system using the distance restraints which have been padded according to Section 2.3.3.

Consider a single assignment for an intermolecular distance restraint with minimum and maximum distances d_l, d_u between atoms $\mathbf{p}, \mathbf{q} \in \mathbb{R}^3$ (see Figure 2.12). Since the restraint must be intermolecular, let $\mathbf{p}$ lie in subunit A and $\mathbf{q}$ lie in subunit B. If we assume the position and orientation of only subunit A are known, then $\mathbf{p}$ is known, but $\mathbf{q}$ is unknown. Let $\mathbf{q}_A$ be the position of the symmetric partner of $\mathbf{q}$ in subunit A. Due to the symmetry, $\mathbf{q}$ is related to $\mathbf{q}_A$ by a rotation about the symmetry axis (whose orientation is parallel to $\hat{\mathbf{z}}$, but whose position $\mathbf{t}$ is unknown):

$$\mathbf{q} \;=\; R(\mathbf{q}_A - \mathbf{t}) + \mathbf{t} \tag{2.1}$$

where R denotes a rotation about $\hat{\mathbf{z}}$ by an angle $\alpha = \frac{2\pi}{n}$ and n is the oligomeric number of the protein. Therefore, to compute positions of the symmetry axis whose oligomer structures satisfy the distance restraint assignment, DISCO computes values of $\mathbf{t}$ such that distance restraint is satisfied: $d_l \leq |R(\mathbf{q}_A - \mathbf{t}) + \mathbf{t} - \mathbf{p}| \leq d_u$.

Since we chose a coordinate system in which the symmetry axis is parallel to $\hat{\mathbf{z}}$, we can simplify this problem to two dimensions instead of three. Construct a plane P perpendicular to $\hat{\mathbf{z}}$ such that it contains $\mathbf{q}$ and $\mathbf{q}_A$. Let $A_3(\mathbf{p}, d_l, d_u)$ be a three-dimensional annulus centered at $\mathbf{p}$ whose radii d_l, d_u are equal to the lower and upper distance bounds of the distance restraint. The intersection of P with $A_3(\mathbf{p}, d_l, d_u)$ yields a two-dimensional annulus $A_2(\mathbf{p}', r_l, r_u)$ where $\mathbf{p}'$ is the projection of $\mathbf{p}$ along $\hat{\mathbf{z}}$ onto P and the radii are: $r_l = \sqrt{d_l^2 - |\mathbf{p} - \mathbf{p}'|^2}$ and $r_u = \sqrt{d_u^2 - |\mathbf{p} - \mathbf{p}'|^2}$. Therefore,

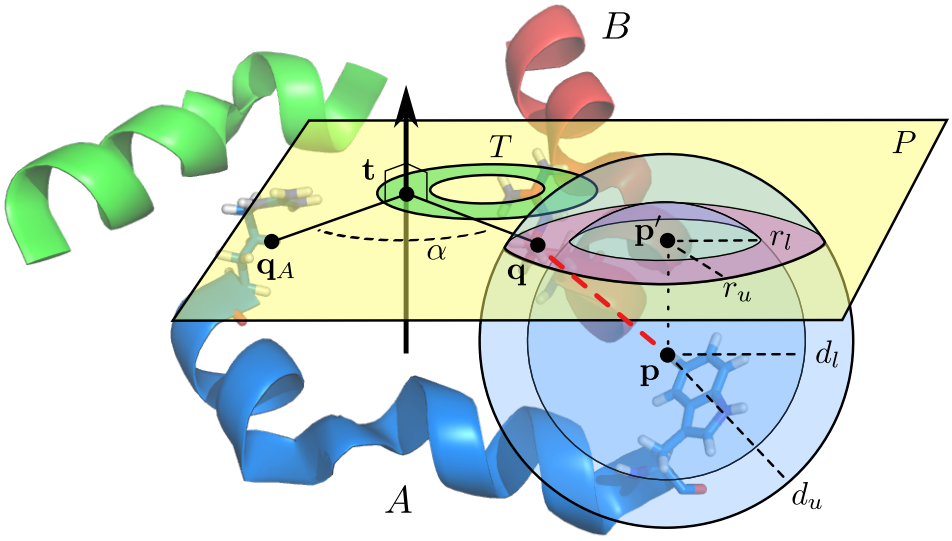


FIGURE 2.12: Symmetric distance restraint geometry for a hypothetical trimer: An inter-molecular distance restraint (red dashed line) between atom $\mathbf{p}$ (in subunit A, whose position is known) and atom $\mathbf{q}$ (in subunit B, whose position is unknown) is satisfied when $\mathbf{q}$ lies between two 3D spheres with radii d_l and d_u centered at $\mathbf{p}$. The orientation of the symmetry axis (black arrow), $\mathbf{q}$, and $\mathbf{q}_A$ (the symmetric partner of $\mathbf{q}$ in subunit A) define a plane P which allows us to reduce the problem to two dimensions: The distance restraint is satisfied when $\mathbf{q}$ lies between two circles with radii r_l and r_u centered at $\mathbf{p}'$, which is the projection of $\mathbf{p}$ onto P along the direction of the symmetry axis. The position of the symmetry axis $\mathbf{t}$ relates $\mathbf{q}$ to $\mathbf{q}_A$ by a fixed angle α. Positions of $\mathbf{t}$ that satisfy the distance restraint compose T, the green annulus in P.

the distance restraint is satisfied when

$$\mathbf{q} \quad \in \quad A_2(\mathbf{p}', r_l, r_u). \tag{2.2}$$

By substituting Eq. (2.1) into Eq. (2.2), we relate the symmetry axis position $\mathbf{t}$ to satisfying positions of $\mathbf{q}$:

$$R(\mathbf{q}_A - \mathbf{t}) + \mathbf{t} \quad \in \quad A_2(\mathbf{p}', r_l, r_u). \tag{2.3}$$

To solve for $\mathbf{t}$, we return to Eq. (2.1) which can be rewritten: $(R - I)\mathbf{t} = R\mathbf{q}_A - \mathbf{q}$. Next, we substitute Eq. (2.2) for $\mathbf{q}$ and lift the operators to set operators to consider set membership in place of strict equality: $(R - I)\mathbf{t} \in R\mathbf{q}_A \ominus A_2(\mathbf{p}', r_l, r_u)$ where $\ominus$ represents the Minkowski difference (Lozano-Perez, 1981). We evaluate the

Minkowski difference by simply translating the annulus:

$$(R - I)\mathbf{t} \quad \in \quad A_2(R\mathbf{q}_A - \mathbf{p}', r_l, r_u). \tag{2.4}$$

Consider all solutions to Eq. (2.4) for $\mathbf{t}$ as a set T, which represents the set of symmetry axis positions whose oligomer structures satisfy the distance restraint assignment:

$$T \quad = \quad \{\mathbf{t} \in \mathbb{R}^2 \mid (R - I)\mathbf{t} \in A_2(R\mathbf{q}_A - \mathbf{p}', r_l, r_u)\}. \tag{2.5}$$

To analyze T, we first note that the matrix $(R - I)$ is the composition of a 2D rotation W and a scaling h. To describe h and W, the 2D rotation matrix R can be expressed as a matrix with two orthogonal column vectors of unit length:

$$R \quad = \quad [\, \mathbf{u} \;\; \mathbf{v} \,]. \tag{2.6}$$

Similarly, $(R - I)$ can be expressed as

$$(R - I) \quad = \quad [\, \mathbf{u} - \hat{\mathbf{x}} \;\; \mathbf{v} - \hat{\mathbf{y}} \,] \tag{2.7}$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the 2D unit axes. Since $\hat{\mathbf{y}} = R_{\frac{\pi}{2}}\hat{\mathbf{x}}$ and $\mathbf{v} = R_{\frac{\pi}{2}}\mathbf{u}$ where $R_{\frac{\pi}{2}}$ is a rotation in the plane of $\frac{\pi}{2}$ radians, then the following must also be true: $\mathbf{v} - \hat{\mathbf{y}} = R_{\frac{\pi}{2}}(\mathbf{u} - \hat{\mathbf{x}})$, thus showing that $\mathbf{v} - \hat{\mathbf{y}}$ and $\mathbf{u} - \hat{\mathbf{x}}$ are orthogonal and right-handed. $\mathbf{u} - \hat{\mathbf{x}}$ and $\mathbf{v} - \hat{\mathbf{y}}$ are not of unit length, but share a common scaling h that normalizes them:

$$\frac{1}{h}|\mathbf{u} - \hat{\mathbf{x}}| \quad = \quad 1 \tag{2.8}$$

$$\frac{1}{h}|\mathbf{v} - \hat{\mathbf{y}}| \quad = \quad 1 \tag{2.9}$$

$$h \quad = \quad |\mathbf{u} - \hat{\mathbf{x}}| = |\mathbf{v} - \hat{\mathbf{y}}|. \tag{2.10}$$

Together, $\mathbf{u} - \hat{\mathbf{x}}$ and $\mathbf{v} - \hat{\mathbf{y}}$ form the basis for the rotation W

$$W \quad = \quad \frac{1}{h}[\, \mathbf{u} - \hat{\mathbf{x}} \;\; \mathbf{v} - \hat{\mathbf{y}} \,]. \tag{2.11}$$

Using h, W, and Eq. (2.5), we can rewrite Eq. (2.4):

$$hWT = A_2(R\mathbf{q}_A - \mathbf{p}', r_l, r_u). \tag{2.12}$$

Since T represents a set and hW is invertible, we have replaced the set inclusion of Eq. (2.4) with strict equality. Solving for T, we see it must also be an annulus in two dimensions:

$$T = A_2\left(\frac{1}{h}W^{-1}(R\mathbf{q}_A - \mathbf{p}'), \frac{r_l}{h}, \frac{r_u}{h}\right). \tag{2.13}$$

Therefore, DISCO computes the annulus T for a single distance restraint assignment exactly and in closed form using Eq. (2.13). If $\mathcal{D} = \{D_i\}$ is the set of distance restraints where $D_i = (\mathbf{p}_i, \mathbf{q}_i)$, DISCO evaluates Eq. (2.13) for each i to compute a set of annuli $\mathcal{T} = \{T_1, \ldots, T_{|\mathcal{D}|}\}$ that lies on P. In the cases where $T_i = \emptyset$ (when $A_3(\mathbf{p}_i, d_l, d_u)$ and P do not intersect), the restraint cannot be satisfied by any oriented oligomer structure. Effectively, $T_i = \emptyset$ indicates the corresponding restraint is inconsistent with respect to the grid orientation and the symmetry.

To account for possible subunit assignments for a distance restraint (e.g., subunit ambiguity), DISCO computes an annulus for each possible subunit assignment of $\mathbf{q}$ by varying the angle of rotation described by R in Eq. (2.13) to choose different subunits. Since the restraint could be interpreted with any one of these possible assignments, and all of them are mutually exclusive, we conservatively encode the choices using a logical OR operator to avoid a combinatorial enumeration of assignment possibilities. Hence, we redefine T_i to encode the annuli from the possible subunit assignments using set union:

$$T_i = \bigcup_{j=1}^{n-1} A_2\left(\frac{1}{h}W^{-1}(R_j\mathbf{q}_A - \mathbf{p}'), \frac{r_l}{h}, \frac{r_u}{h}\right) \tag{2.14}$$

where R_j is a rotation about the $\hat{\mathbf{z}}$ axis by an angle of $j\alpha$. T_i now represents the set

of symmetry axis positions that satisfy at least one possible subunit assignment for the distance restraint.

Atom ambiguity, which characterizes a distance restraint that could be assigned to multiple pairs of atoms, often due to overlapping chemical shifts, can also be encoded using a union of annuli. We now redefine D_i to represent set of possible atom assignments $\{(\mathbf{p}_k, \mathbf{q}_k)\}$ where $\mathbf{p}_k$ and $\mathbf{q}_k$ are the two atoms for assignment k. Then, T_i can be defined as the union of annuli resulting from all possible atom assignments for the distance restraint:

$$T_i = \bigcup_k A_2\left(\frac{1}{h}W^{-1}(R\mathbf{q}_A^{(k)} - \mathbf{p}'_k), \frac{r_l}{h}, \frac{r_u}{h}\right) \tag{2.15}$$

where $\mathbf{q}_A^{(k)}$ represents the symmetric partner of $\mathbf{q}_k$ in subunit A. Whether a distance restraint possesses atom ambiguity or subunit ambiguity, DISCO represents the set of satisfying symmetry axis positions as a union of annuli in P.

For the annulus analysis to be meaningful, we require the distance restraints to be strictly intermolecular. If a distance restraint were to possess possible intramolecular assignments, then it is possible for the true assignment of the distance restraint to be strictly intramolecular. Since an intramolecular distance restraint cannot possibly characterize the oligomeric structure of the protein, no annulus can be computed for an intramolecular assignment. Hence, the remaining intermolecular assignments must all be incorrect, resulting in an incorrect union of annuli. If the incorrect unions of annuli outnumber the correct unions of annuli, and they all happen to conspire and share some common region, then the resulting computed MSRs may not correctly describe the oligomer structure. Therefore, distance restraints that have possible intramolecular assignments, such as PREs, must not be used, unless other reasoning or data can rule out their intramolecular interpretations.

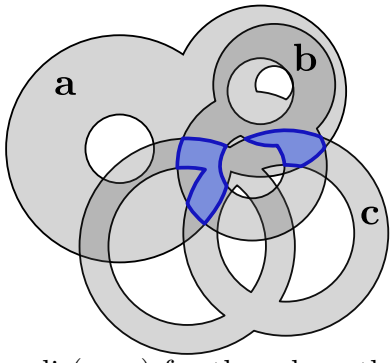


FIGURE 2.13: Unions of annuli (grey) for three hypothetical distance restraints (a, b, and c): Computing the arrangement of the unions of annuli gives all intersection points of the circles bounding the annuli, all edges between intersection points, and all faces bounded by the edges. This example shows two MSRs (blue) which are the two faces of the arrangement contained in all three unions of annuli.

2.3.5 Computing MSRs

To compute positions of the symmetry axis whose oligomer structures satisfy the maximal number of distance restraints (the MSRs), DISCO simultaneously evaluates the unions of annuli for all the distance restraints computed in Section 2.3.4. Ideally, the intersection of all unions of annuli will define a set of symmetry axis positions, but noise and incorrect assignments can result in an empty intersection. Instead, DISCO computes the arrangement of the unions of annuli (see Figure 2.13 for explanation) using the CGAL software library (Hanniel and Halperin, 2000) and chooses as MSRs the faces from the arrangement that are contained in the greatest number of unions of annuli. CGAL is a C++ software library that implements algorithms from computational geometry (such as computing arrangements) and guarantees exact numerical precision. Further details of our algorithm and an analysis of its asymptotic complexity are presented in (Martin et al., 2011b). While the arrangement can, in theory, contain multiple faces with equal restraint satisfaction, the computational tests for DAGK (Figure 2.5) and the GB1 domain-swapped dimer (Figure 2.8) yielded only a single simply-connected MSR in each case.

2.3.6 Computing discrete oligomer structures

The MSRs computed in Section 2.3.5 define continuous sets of symmetry axis positions and consequently describe continuous sets of oligomer structures which are difficult to analyze directly. Therefore, to perform detailed structural analyses of the oligomer structures described by the MSRs, DISCO samples discrete symmetry axis positions from the MSRs on a uniform grid at a user-specified resolution. The axis position sampling is repeated for each grid orientation resulting in a set of complete symmetry axes that vary by orientation as well as position. The sampled axes define rigid transformations that, when applied to the subunit structure, generate symmetric oligomer structures. Figure 2.2 illustrates an example using a trimer.

2.3.7 Evaluating computed structures

DISCO evaluates each computed structure for restraint satisfaction and intermolecular packing, as measured by van der Waals energy. DISCO performs local structure energy-minimization using X-PLOR (Schwieters et al., 2003) where oligomer structures are refined with 1000 steps of Cartesian minimization. The backbone is completely fixed, but the side chains are allowed to move. They are restrained by a van der Waals potential (with default parameters), an NOE potential (with a weight of 30), and the default chemical potentials: `BOND`, `ANGL`, and `IMPR`. Inconsistent distance restraints can optionally be excluded from the the NOE potential. After minimization, DISCO computes the distance restraint satisfaction score by evaluating the RMSD of the distance restraints using their (un-padded) upper distances. DISCO also computes the van der Waals packing score for each minimized structure using the pairwise Lennard-Jones potential. The structural ensemble returned by DISCO is composed of the minimized structures. Finally, since we expect an oligomer structure to have better packing than the subunit in isolation, structures with van der Waals energies higher than the subunit structure are removed from the final

computed ensemble.

2.3.8 NOE atom ambiguity simulation

We simulated atom ambiguity for the NOEs for the GB1 domain-swapped dimer (Byeon et al., 2003) by expanding the assignments to include protons with similar chemical shifts. Using the ^{1}H, ^{13}C, and ^{15}N chemical shifts deposited in the BMRB (Ulrich et al., 2007) (1Q10: 5875) along with the intermolecular NOEs deposited in the PDB (Berman et al., 2000) (1Q10), we simulated two 3D X-filtered NOESY experiments: 3D ^{15}N-edited-HSQC-NOESY and 3D ^{13}C-edited-HSQC-NOESY. We expanded the assignments for an NOE between protons $\mathbf{p}$ and $\mathbf{q}$ in the following way.

Let $H(\mathbf{p})$ be the value of the ^{1}H chemical shift for proton $\mathbf{p}$, and similarly $H(\mathbf{q})$ for $\mathbf{q}$. Let $\Theta(\mathbf{p})$ be the chemical shift of the heavy atom covalently bonded to $\mathbf{p}$. For example, when $\mathbf{p}$ is bonded to a N atom, $\Theta(\mathbf{p})$ is the ^{15}N chemical shift of the bonded N atom. We can view the point $(H(\mathbf{p}), \Theta(\mathbf{p}))$ as residing in a 2D dimensional chemical shift space. In this space, finding protons with similar chemical shifts corresponds to finding *neighbors* of $\mathbf{p}$. We define a proton $\mathbf{s}$ a neighbor of $\mathbf{p}$ if (and only if) the following criteria are satisfied:

$$|H(\mathbf{p}) - H(\mathbf{s})| \leq \delta_H \tag{2.16}$$

$$|\Theta(\mathbf{p}) - \Theta(\mathbf{s})| \leq \delta_\Theta \tag{2.17}$$

where δ_H and δ_Θ are user-specified similarity parameters. Since we are simulating 3D NOESY experiments, we must treat $\mathbf{q}$ as if we do not know $\Theta(\mathbf{q})$. Therefore, DISCO searches for neighbors of $\mathbf{q}$ using only the criterion in Eqn. 2.16. Alternatively, we could interpret $\mathbf{q}$ as having a known $\Theta(\mathbf{q})$ instead of $\mathbf{p}$, but with only chemical shifts, we cannot determine which interpretation was used during assignment. Therefore, we arbitrarily chose $\mathbf{p}$ to have known $\Theta(\mathbf{p})$ for all NOEs.

2.4 Conclusion

DISCO can accurately determine the oligomer structures of proteins with cyclic symmetry using RDCs and distance restraints such as NOEs and disulfide bonds. It provides a graphical analysis of the distance restraints and is able to differentiate between consistent and inconsistent distance restraints using the maximally satisfying regions. Since DAGK and the GB1 domain-swapped dimer are both high-quality solved structures, it is not surprising DISCO did not discover any inconsistent restraints. DISCO's inconsistency analysis is likely to be more useful during earlier stages of structure calculation when distance restraint assignments may be less certain. DISCO computes oligomer structures using intermolecular distance restraints even when precise atom and subunit assignments are not known, thus reducing the need to assign distance restraints unambiguously for structure determination. However, only distance restraints with strictly intermolecular possible assignments must be used. Distance restraints with possible intramolecular assignments (such as PREs) cannot be used without first attempting to discard the distance restraints whose true assignments are intramolecular.

DISCO requires a subunit structure to build models of the oligomeric state, but computing an accurate model of the subunit structure in isolation using intramolecular distance restraints can sometimes be challenging. If intramolecular distance restraints are insufficient to adequately constrain the subunit structure, it may be necessary to record additional RDCs and use an RDC-first approach (Zeng et al., 2009). As an alternative, one could model adjacent subunits during subunit structure calculation using intermolecular restraints to ensure the subunit structure presents an interface amenable to oligomerization. Additionally, for trimers and higher-order oligomers, we expect an alignment tensor computed from the RDCs and the subunit structure will have zero rhombicity. If the rhombicity is significantly greater than

zero, the RDCs do not reflect the oligomeric symmetry, they may not accurately describe the orientation of the symmetry axis, and it will not be possible to apply DISCO. For dimers, the rhombicity is not able to indicate agreement between the symmetry axis orientation and the RDCs, but DISCO is able to search for the best symmetry axis orientation among the three available possibilities; namely, the three eigenvectors of the alignment tensor.

Since DISCO can compute the exact set of oriented oligomer structures that satisfy the distance restraints for each grid orientation, the variance in atom position of the computed ensemble of structures yields a meaningful measure of the range of oligomer structures allowed by the distance restraints. DISCO's graphical analysis is easy to visualize and can find distance restraints that are inconsistent with the RDCs, or are inconsistent with other distance restraints. The entire protocol has been completely automated in a software package that will be freely available and open-source upon publication.

2.5 Appendix: Perturbation analysis of the arrangement

Let $\beta \in [0, 100]$ be a padding percentage applied to the bounds of all distance restraints such that the lower bound is multiplied by

$$\left(1 - \frac{\beta}{100}\right) \tag{2.18}$$

and the upper bound is multiplied by

$$\left(1 + \frac{\beta}{100}\right). \tag{2.19}$$

For this perturbation analysis, the parameter β completely replaces the scheme for padding based on the variance of the subunit ensemble. To explore the effect of the parameter β on structure calculation, we computed MSRs from the disulfide

bonds for DAGK for values of β varying from 0 to 15 in increments of 0.25. For each of the resulting 61 sets of MSRs, we computed the minimum and maximum distances from the MSRs to the position of the symmetry axis of the reference structure. Figure 2.14 (A) shows these distance ranges plotted against β. To estimate the quality of oligomer structures represented by these MSRs, we sampled symmetry axis positions from the MSRs (and hence, oligomer structures) at a very fine resolution (0.0125 Å) and computed their backbone RMSDs to the reference structure for DAGK. The resulting ranges of backbone RMSDs are shown in Figure 2.14 (B). The ranges of reference axis position/MSR distances and the ranges of backbone RMSDs closely resemble each other, indicating that DISCO's geometric analysis is able to accurately represent differences in oligomer structures using differences in the symmetry parameters. Interestingly, even though the MSRs for $\beta \in [1, 5]$ allow for sampling arbitrarily close to the reference symmetry axis position, the minimum achievable backbone RMSD was 0.0490 Å. Since oligomer structures computed by DISCO are symmetric by construction, comparisons with the reference structure for DAGK (which has slight deviations from perfect symmetry) will not yield perfect matches.

In general, the size of the ranges in Figure 2.14 increase with β, but there are three interesting exceptions. With $\beta = 0$, 21/24 of the disulfide bond distance restraints are satisfied by the MSRs. The MSRs computed when $\beta = 5.5$ are able to satisfy an additional distance restraint, increasing the count to 22. The number of satisfied distance restraints increases again at $\beta = 6.25$, and once more at $\beta = 7.5$, where all 24 distance restraints are satisfied by the MSRs. These three values of β, where the number of satisfied distance restraints increases, define four β *intervals* over which the number of satisfied distance restraints remains constant. In the four different β intervals, the geometry of the MSRs (Figure 2.15) is markedly different.

As can be seen from Figure 2.15, the size of the MSRs grow regularly in the first

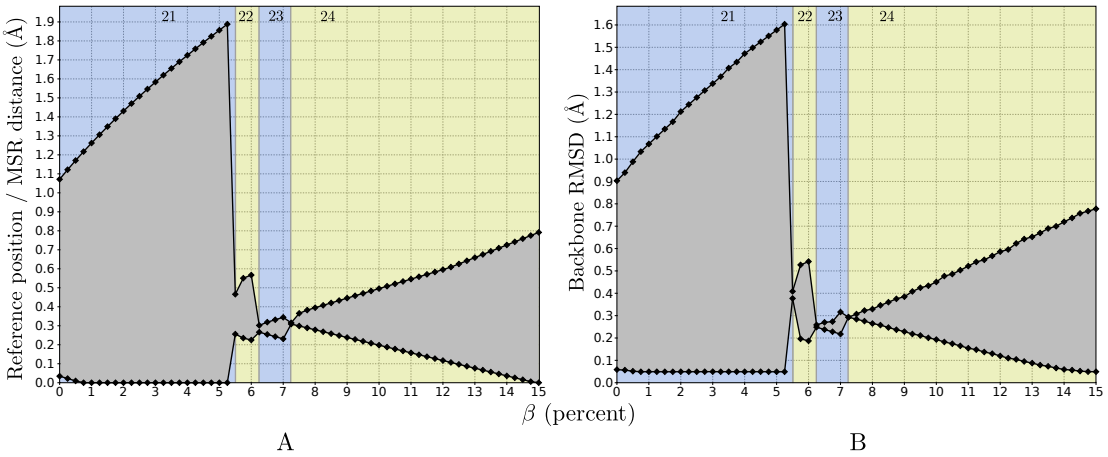


FIGURE 2.14: (A): Range of distances between the MSRs and the reference symmetry axis position for varying values of β. The top series shows the maximum distances and the bottom series shows the minimum distances. The four yellow and blue regions show the intervals of β sharing the same number of satisfied disulfide bond distance restraints. The number of satisfied distance restraints is shown at the top of the interval. (B): Range of backbone RMSDs between the reference structure and oligomer structures sampled very finely from the MSRs.

β interval of $[0, 5.25]$, since the outer radii of the distance restraint annuli also grow regularly with β. However, at $\beta = 5.5$, the MSRs "jump" to a new position, since the arrangement now defines a deeper face corresponding to the satisfaction of the additional disulfide bond distance restraint. As β increases over the next β interval of $[5.5, 6.0]$, the MSRs grow regularly again until $\beta = 6.25$, where the satisfaction of an additional distance restraint causes another "jump." This grow-then-jump pattern continues until all distance restraints are satisfied. Afterwards, the MSRs simply grow regularly with β since there are no more distance restraints to satisfy.

2.6 Appendix: Analysis of the arrangement of unions of annuli

Once a union of annuli is computed for each distance restraint, DISCO computes the arrangement $\mathcal{A}$ of all the circular curves bounding the unions of annuli. $\mathcal{A}$ is computed using a randomized incremental algorithm (Halperin, 1997), implemented in the CGAL library (Hanniel and Halperin, 2000). $\mathcal{A}$ represents all intersection

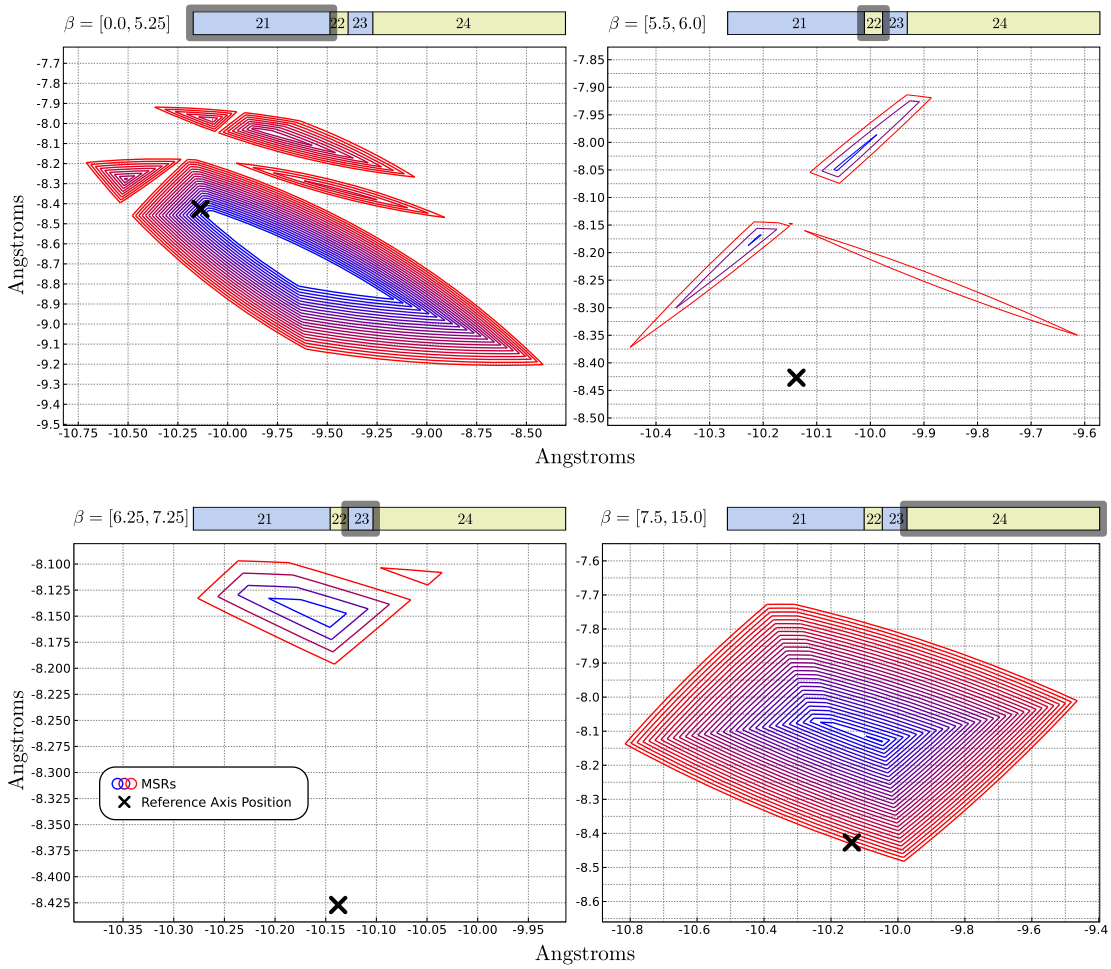


FIGURE 2.15: An overlay of the MSRs computed from the disulfide bonds for DAGK for varying values of β. Each of the four β intervals is shown as a separate plot. The β interval itself is shown at the top of each plot using the yellow/blue colors from Figure 2.14. Within each plot, the reference symmetry axis position is marked with a black X and the MSRs for all the β values of the interval are shown together as curves using a gradient of colors, from blue to red as β increases.

points of the circular curves, all edges bounded by the intersection points, and all faces bounded by the edges. We refer the faces of $\mathcal{A}$ contained in the greatest number of unions of annuli as the *maximally satisfying regions* (MSRs). These faces represent symmetry axis positions that satisfy the greatest number of inter-subunit distance restraints.

Formally, let each face f in $\mathcal{A}$ have an associated *depth*, $d(f)$, equal to the number of unions of annuli that contain f. The MSRs are the faces in $\mathcal{A}$ with the maximum

depth, and the unbounded face f_u has a depth of zero. The MSRs are found by analyzing the dual graph $G = (V, E)$ of $\mathcal{A}$ (Figure 2.16) where V contains one node for each face in $\mathcal{A}$, including f_u. Let f' represent the dual of f, which is the vertex in V corresponding to f. E contains an edge (f'_1, f'_2) when two faces f_1 and f_2 in $\mathcal{A}$ share an edge. To annotate the faces of $\mathcal{A}$ with depths, DISCO performs breadth-first search (BFS) beginning at the vertex f'_u. When BFS traverses an edge in E, this corresponds to crossing an edge h from the previous face f_p to the next face f_n in $\mathcal{A}$. Therefore, $d(f_n)$ is assigned $d(f_p) - 1$ if crossing h leaves a union of annuli, $d(f_p) + 1$ if crossing h enters a union of annuli, or $d(f_p)$ if h lies in the interior of a union of annuli (i.e., the depth remains the same, see Figure 2.17). Once the faces of $\mathcal{A}$ have been annotated with depths, DISCO returns the MSRs by enumerating the faces with maximum depth.

To construct $\mathcal{A}$, the circles bounding the unions of annuli are decomposed into x-monotone circular arcs, which are restricted to be monotonic in the x-direction (i.e., no vertical line intersects the curve more than once). The two x-monotone circular arcs resulting from the circle decomposition are subsequently divided into smaller arcs resulting from intersections with other x-monotone circular arcs during construction of $\mathcal{A}$. We will say the resulting circular arcs are all *supported* by the original circle. To decide whether a crossing enters/leaves or remains within a union of annuli, we determine if the edge h lies in the interior of the union of annuli whose constituent circles support h. Since each edge in $\mathcal{A}$ can only be supported by a circle from a single union of annuli, the supporting union of annuli can be referenced by a pointer stored at the edge, and this pointer can be set during the construction of the arrangement. If the midpoint of the edge h lies in the interior of its union of annuli, then h is an *interior edge* and the crossing remains within the union of annuli. Any point along h (except for the endpoints) can be used to test if h is an interior edge, but the midpoint is the most numerically-stable choice. If h lies on the boundary of

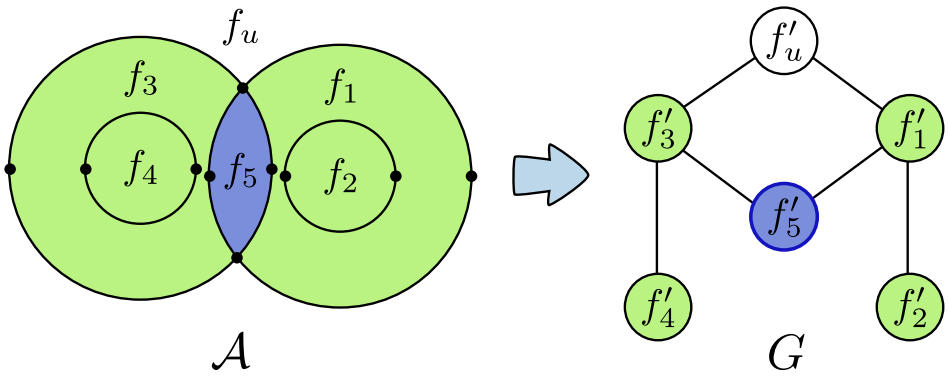


FIGURE 2.16: A sample dual graph (G) for a hypothetical arrangement ($\mathcal{A}$) of x-monotone curves bounding two annuli, showing the single MSR (blue), the remaining bounded faces (green), and the unbounded face (f_u). The bounded faces in $\mathcal{A}$, which are labeled $f_1, \cdots, f_5$, map to vertices in G, which are labeled $f'_1, \cdots, f'_5$. The unbounded face maps to the vertex f'_u.

its union of annuli (i.e., is not an interior edge), the crossing enters or leaves a union of annuli. The orientation of the circle supporting h is used to encode whether or not the circle defines the interior or exterior boundary of the union of annuli.

2.7 Appendix: Analysis of complexity

In this appendix, we prove bounds on the time and space complexity of the computation of the MSRs from ambiguously-assigned inter-subunit distance restraints.

Lemma 1. *For an oligomeric protein complex with cyclic symmetry and n distance restraints assigned ambiguously, each having at most s possible assignments, the MSRs can be computed in expected $O(s^3n^2)$ time and $O(s^2n^2)$ space.*

Proof. DISCO computes a single annulus for each assignment of each distance restraint which results in n unions of annuli, each having at most s annuli. Hence, there are sn annuli in total and each union of annuli has a complexity of $O(s)$. In the next step, DISCO decomposes the boundaries of the unions of annuli into x-monotone circular arcs, four for each annulus, resulting in $O(sn)$ circular arcs. The computation of the arrangement $\mathcal{A}$ can be accomplished using a randomized incremental algorithm

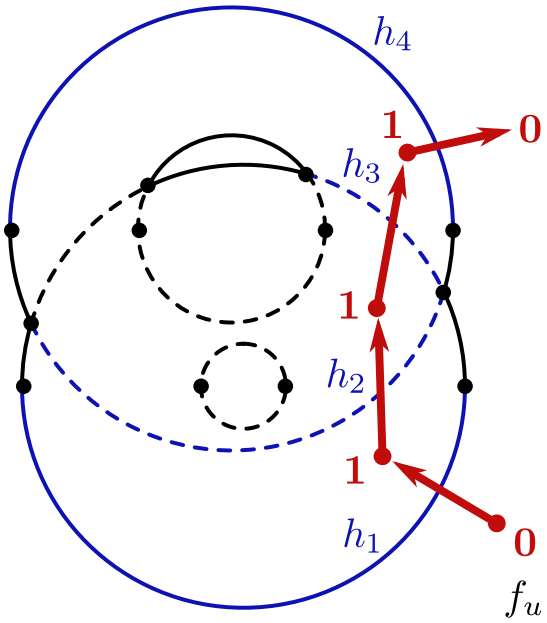


FIGURE 2.17: A path (red arrows) visits faces in a union of annuli (blue/black curves) starting with the unbounded face (f_u) and crosses four edges: h_1, h_2, h_3, and h_4 (blue curves). Interior edges of the union of annuli are shown with dashed curves. Since f_u is initialized with a depth of zero, crossing h_1 increases the depth to one, crossing the interior edges h_2 and h_3 do not modify the depth, and crossing h_4 returns the depth to zero.

requiring expected $O(sn\log(sn)+k)$ time and $O(s^2n^2)$ space, where k is the number of intersections in the arrangement (Halperin, 1997). The depths can be stored using constant space at each face, thus leaving the $O(s^2n^2)$ space requirements of the original algorithm unchanged. To test if an edge h lies on the interior of its supporting union of annuli, we must determine whether its midpoint lies in the interior. Since the single union of annuli supporting h can be found in constant time by following the pointer at h, and the complexity of each union of annuli is $O(s)$, the interior predicate for h can be evaluated in $O(s)$ time. The complexity of $\mathcal{A}$ is bounded by $O(s^2n^2)$, so the dual graph G has $O(s^2n^2)$ nodes and $O(s^2n^2)$ edges, hence the interior predicate will be evaluated $O(s^2n^2)$ times. Therefore, BFS on G can be performed in $O(s^3n^2)$ time using $O(s^2n^2)$ space.

Finally, to find the MSRs, the faces of $\mathcal{A}$ (which have been annotated with depths)

can be enumerated in $O(f)$ time, where f is the number of faces in the arrangement. Thus, the time required to compute the MSRs from n distance restraints, each having at most s possible assignments, is expected $O(sn \log(sn) + k) + O(s^3n^2) + O(f)$ and is output-sensitive. Since the total complexity of the arrangement is bounded by $O(s^2n^2)$, $f + k$ is also bounded by $O(s^2n^2)$, and therefore we can simplify the total time to expected $O(s^3n^2)$. The overall space requirements depend on the size of $\mathcal{A}$, which is $O(s^2n^2)$, and the size of G, which is $O(s^2n^2)$. Therefore, the total space required is $O(s^2n^2)$ as well. □

Next, we will use biophysical facts to place bounds on the number of possible assignments for an inter-subunit distance restraint and simplify the complexity bounds.

Lemma 2. *For an oligomeric protein complex with cyclic symmetry and n distance restraints assigned with subunit and/or atom ambiguity, the MSRs can be computed in expected $O(n^2)$ time and $O(n^2)$ space.*

Proof. For subunit ambiguity, the number of possible assignments is bounded by the number of subunits in the complex. In proteins for *E. coli*, only 2.2% of proteins annotated with subunit designations are composed of more than 12 subunits (Goodsell and Olson, 2000). Therefore, we assume the oligomeric number of protein complexes is bounded by a constant. For atom ambiguity, the number of possible assignments is bounded by the spectral overlap of neighboring resonances and peaks in NMR spectra. In practice, for proteins of up to around 200 residues, 3D NOESY experiments (Marion et al., 1989) are sufficient to limit spectral overlap to a constant amount per peak. For larger proteins, 4D NOESY (Kay et al., 1990), which uses an extra dimension to resolve cross peaks (similar in ways to a lifting transform), may be required to limit spectral overlap to a constant amount. Even higher-dimensional NMR experiments are possible (Kim and Szyperski, 2003). Since the oligomeric number of protein complexes and the amount of spectral overlap per peak can be

bounded by a constant, the number of possible assignments for a distance restraint assigned with subunit and/or atom ambiguity is also bounded by a constant. Consequently, each union of annuli has a constant number of annuli. Therefore, s is $O(1)$ and the bounds of Lemma 1 simplify to expected $O(n^2)$ time and $O(n^2)$ space. $\square$

3

Structure of an HIV-1 neutralizing antibody target, the lipid-bound gp41 envelope membrane proximal region trimer

The text of this chapter has been adapted from a published manuscript that was co-authored with Patrick N. Reardon, Harvey Sage, S. Moses Dennison, Bruce R. Donald, S. Munir Alam, Barton F. Haynes, and Leonard D. Spicer. In this section, my primary contribution was applying DISCO *to confirm the trimeric structure of gp41-M-MAT and assisting with the analysis of NMR data.*

P. N. Reardon, H. Sage, S. M. Dennison, J. W. Martin, B. R. Donald, S. M. Alam, B. F. Haynes, and L. D. Spicer. Structure of an HIV-1 neutralizing antibody target, the lipid-bound gp41 envelope membrane proximal region trimer. Proceedings of the National Academy of Sciences, 2014. 111(4):1391–1396.

Abstract: The membrane proximal external region (MPER) of HIV-1 gp41 is involved in viral-host cell membrane fusion. It contains short amino acid sequences that are binding sites for the HIV-1 broadly neutralizing antibodies 2F5, 4E10 and 10E8, making these binding sites important targets for HIV-1 vaccine development.

We report a high-resolution structure of a designed MPER trimer assembled on a detergent micelle. The NMR solution structure of this trimeric domain, designated gp41-M-MAT, shows the three MPER peptides each adopt symmetric α-helical conformations exposing the amino acid side chains of the antibody binding sites. The helices are closely associated at their N-termini, bend between the 2F5 and 4E10 epitopes and gradually separate toward the C-termini where they associate with the membrane. Monoclonal antibodies 2F5 and 4E10 bind gp41-M-MAT with nanomolar affinities, consistent with the substantial exposure of their respective epitopes in the trimer structure. The traditional structure determination of gp41-M-MAT using the Xplor-NIH protocol was validated by independently determining the structure using the DISCO sparse-data protocol, which exploits geometric arrangement algorithms that guarantee to compute all structures and assignments that satisfy the data.

Significance: A major roadblock in the development of an HIV vaccine is the need to develop vaccine regimens that will induce antibodies that bind to conserved regions of the HIV envelope and neutralize many different virus quasispecies. One such envelope target is at the region closest to the membrane, the gp41 membrane proximal external region (MPER). Previous work has demonstrated that antibodies that target this region bind both to the gp41 polypeptide and to the adjacent viral membrane. However, what has been missing is a view of what the MPER neutralizing epitopes may look like in the context of a trimeric orientation with lipids. We have constructed an MPER trimer associated with lipids and solved the trimer structure by NMR spectroscopy.

3.1 Introduction

Infection of a CD4+ T-cell by HIV-1 is mediated by the envelope protein (Env), a trimeric complex located on the virion surface that consists of three copies each of gp120 and gp41. This complex is a macromolecular machine responsible for host cell recognition followed by fusion of the viral and CD4+ T cell membranes, leading to virus entry (Freed, 2001). The Env complex represents the primary target for antibody-mediated viral neutralization (Burton et al., 2004).

The Env protein complex undergoes dramatic conformational changes during the process of membrane fusion. Biochemical and structural evidence suggests that membrane fusion involves at least three states of the Env complex (Gallo et al., 2003; Harrison, 2008). The first state is the resting pre-fusion state that exists prior to host cell encounter and receptor binding. This state has been studied by several groups using cryo-EM (Zhu et al., 2006; Zanetti et al., 2006; White et al., 2010; Wu et al., 2010; Bartesaghi et al., 2013; Mao et al., 2012). The second state is a pre-fusion intermediate where gp41 is interacting with both the host cell and viral membranes. This pre-fusion intermediate, or a closely related intermediate, is also believed to be the target for fusion inhibiting peptides (Ashkenazi and Shai, 2011) as well as the broadly neutralizing antibodies, 2F5 and 4E10 (Frey et al., 2008). The final state is the post-fusion or six-helix bundle. The formation of this conformation is thought to drive membrane fusion. This conformation is stable, and its structure has been well studied using X-ray crystallography techniques (Buzon et al., 2010). Binding studies have shown that the broadly neutralizing antibodies 2F5 and 4E10 do not bind with high affinity to either the post-fusion six helix bundle or the pre-fusion resting state, suggesting that a pre-fusion intermediate state is the target for these antibodies (Frey et al., 2008).

The membrane proximal external region (MPER) is a 28-residue segment of each

subunit in the gp41 homotrimer. This tryptophan-rich segment is juxtaposed to the transmembrane domain and plays an important role in the membrane fusion process leading to viral infection of the host cell (Muñoz-Barroso et al., 1999; Salzwedel et al., 1999). The MPER contains the recognition sites (binding *epitopes*) for several broadly neutralizaing antibodies including 2F5, 10E8 and 4E10 (Muster et al., 1993; Stiegler et al., 2001; Huang et al., 2012). This has motivated efforts to develop vaccines designed to induce antibodies specific to this region. Vaccine candidates based on linear peptides from the MPER (Alam et al., 2008), trimeric gp41 constructs (Hinz et al., 2009; Lenz et al., 2005), and conformationally constrained peptides have been previously reported (Guenaga et al., 2011; Ofek et al., 2010a). In animal models, many of these vaccine designs have elicited antibodies that recognize epitopes in the MPER (Alam et al., 2008; Guenaga et al., 2011; Ofek et al., 2010a). However, none of the induced plasma antibodies strongly neutralize HIV-1, (Alam et al., 2008; Hinz et al., 2009; Ofek et al., 2010a; Burton, 2010) either because the trial vaccines do not present the epitope residues in a native conformation or in the presence of the correct molecular environment, or because of limitation of induction of MPER antibodies by host tolerance mechanisms (Verkoczy et al., 2010, 2013; Chen et al., 2013; Doyle-Cooper et al., 2013).

Monoclonal antibodies (Mabs) 2F5 and 4E10 are polyreactive for non-HIV-1 proteins and for lipids (Haynes et al., 2005; Yang et al., 2013). Crystal structures of 2F5 and 4E10 antigen-binding fragment (Fab) domains bound to short epitope-containing MPER peptides show limited CDR-H3 contacts with the MPER peptides, and together with the lipid-reactive data, prompted speculation that the long hydrophobic CDR-H3 loops in the antibodies contact the viral membrane (Ofek et al., 2004; Cardoso et al., 2005; Phogat et al., 2008). Mutations of some of the hydrophobic residues in the CDR-H3 regions reduce the lipid binding activity of these antibodies without reducing peptide binding, but these mutants are also non-neutralizing for HIV-1

(Alam et al., 2009; Ofek et al., 2010b; Scherer et al., 2010). These data demonstrated that association with the viral membrane plays an important role in the molecular mechanism for viral neutralization by 2F5 and 4E10 (Alam et al., 2009).

Here we report the biosynthesis and structure determination of a micelle-bound MPER trimer in a putative prefusion intermediate state. The designed trimer readily associates with dodecylphosphocoline (DPC) micelles and 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) liposomes. The solution NMR structure reported is an atomic level representation of the MPER trimer that is complexed directly with the outer surface of the micelle. This trimer, displayed on both liposomes and micelles, avidly binds the 2F5 and 4E10 neutralizing antibodies. We also observe conformational flexibility within the polypeptide subunits that we hypothesize is important for binding to the 2F5 and 4E10 antibodies based on the crystal structures of the antibodies bound to short peptide epitopes.

3.2 Results

3.2.1 Trimer MPER Construct and NMR Structure

We designed the gp41 MPER trimer construct based on a chimeric polypeptide monomer containing the 27-residue trimerization domain from bacteriophage T4 fibritin (the foldon domain) N-terminal to the MPER sequence

`NEQELLELDKWASLWNWFNITNWLWYIK`

corresponding to residues 656-683 of the Env protein from the Hxb2 strain of HIV-1. The foldon domain is linked to the MPER domain via a flexible Gly-Ser-Ser-Gly linker, allowing the MPER to adopt orientations with minimal constraint from the foldon. The polypeptide spontaneously trimerizes and directly associates with the phospholipid membrane surface at the MPER C-terminus, where the transmembrane segment of gp41 begins in the full-length protein. This transmembrane domain

FIGURE 3.1: gp41-M-MAT design and purification. A. The gp41-M-MAT peptide is expressed as a TrpLE fusion construct in C41(DE3) *E. coli* cells. The TrpLE tag directs the polypeptide into inclusion bodies which facilitates expression and purification. B. gp41-M-MAT after cyanogen bromide cleavage to remove the TrpLE tag. C. Peptide sequence of MPER used in the gp41-M-MAT. The MPER residues are numbered 32–59 in gp41-M-MAT, which corresponds to residues 656–683 of Env protein in the HXB2 strain of HIV-1. Epitopes for 2F5 and 4E10 are shown in blue and red, respectively.

consists of 31 residues that are not included in our construct. A schematic diagram of the gp41 MPER containing Membrane Associated Trimer designated gp41-M-MAT is shown in Figure 3.1.

The solution structure of the membrane-associated gp41-M-MAT was determined using NOE-based distance restraints, dihedral angles, 3J-coupling constants, and HN residual dipolar coupling data collected on NMR spectrometers at 600, 800 and 950 MHz. The ^{15}N-TROSY HSQC spectrum revealed a single set of amide moiety resonances, which facilitated the assignments and indicated that the trimer was symmetric. The solution structure we determined for gp41-M-MAT is shown in Figure 3.2. A comparison of the structures computed by Xplor-NIH and the alternative DISCO method is shown in Figure 3.3 and Table 3.1. Importantly, these two alternative methods result in nearly identical trimeric architectures for the MPER. The most significant differences in the structural ensembles derived by the two methods are in the flexible linkers. These differences lead to increased RMSDs between the two gp41-M-MAT ensembles since the conformational variations in the linkers allow in-

Table 3.1: Structural comparison between the ensembles computed by Xplor-NIH and DISCO.

Region	Residues	RMSD to the mean[1]		RMSD between means[2]
		Xplor-NIH	DISCO	
gp41-M-MAT	1–59	1.54	1.68	2.38
Foldon	1–27	0.75	0.34	2.06
Linker	28–31	1.26	1.77	2.04
MPER	32–59	1.42	1.57	0.98

For each region, all comparisons were made under optimal alignment and all RMSDs are reported in Å. [1]Average backbone RMSD to the mean structure. [2]For each region, the RMSD between the mean structure from the Xplor-NIH and DISCO based ensembles is reported.

creased rotational freedom around the symmetry axis that results in a small range of relative rotations for the foldon and MPER domains. The ensemble of gp41-M-MAT structures computed by Xplor-NIH and the ensemble computed by DISCO have been deposited to the Protein Data Bank (PDB) Berman et al. (2000) under the codes 2LP7 and 2M7W respectively.

Based on the HN residual dipolar coupling (RDC) data, the calculated rhombicity of the trimer on the micelle was 0.052. This was close to zero, consistent with a symmetric homotrimer, where the ideal rhombicity is zero. Further analysis of this data revealed that the separate foldon and MPER domains adopted the same symmetry axis orientations, within experimental error. A summary of the symmetry axis analysis is shown in Figure 3.4 and the agreement of the subunit structure to the RDC data is shown in Figure 3.5.

Each subunit of the MPER trimer is in an α-helical conformation approximately 40 Å long. The three helices form a three-fold symmetric left-handed bundle that progressively expands from residue 32 where the C^{α} atoms of each helix are separated by 14 Å to a C-terminal separation of ~30 Å between C^{α} positions in residue 59. Intermolecular NOEs are observed between the side chains of residues 39 and 42 and the backbone amide of residue 38. Additional intermolecular NOEs were not observed

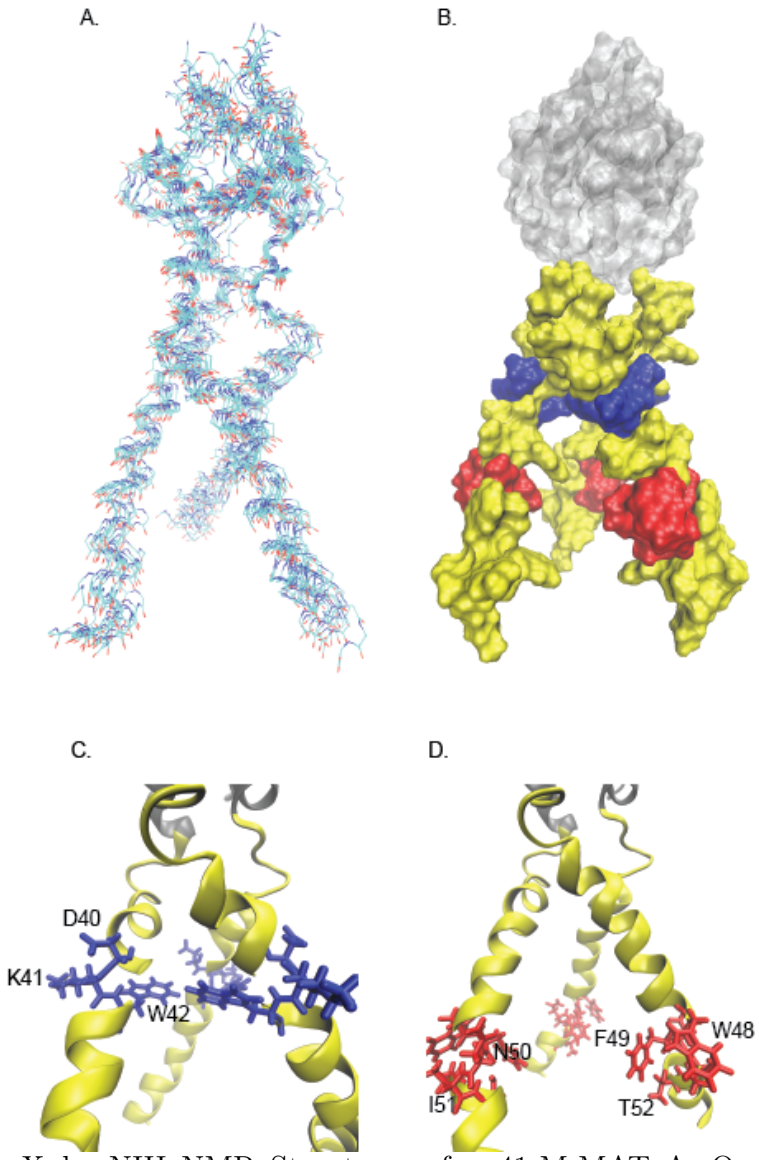
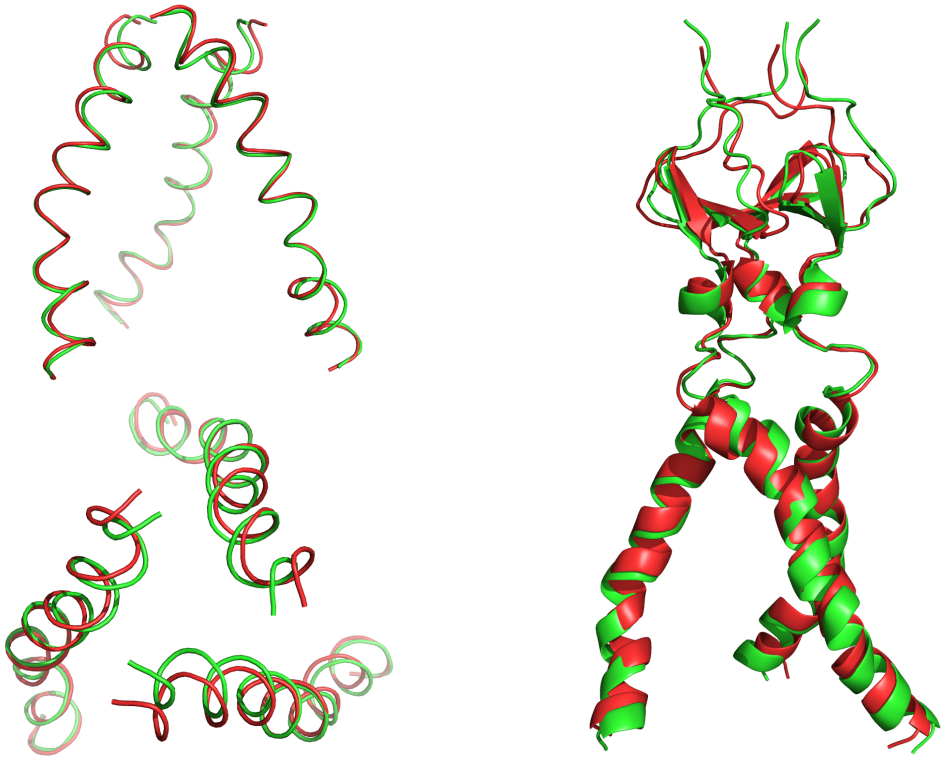


FIGURE 3.2: Xplor-NIH NMR Structures of gp41-M-MAT A. Overlay of 11 low energy gp41-M-MAT structures containing no NOE violations or dihedral violations of greater than 0.5 Å or 5° respectively. B. Space-filling diagram of the minimized average structure with the 2F5 and 4E10 epitopes colored in blue and red respectively. The Foldon domain is shown in translucent light gray. C and D. Detailed views emphasizing the epitopes for 2F5 in blue (C) and 4E10 in red (D).



(a) MPER: 1.07 Å backbone RMSD (b) gp41-M-MAT: 1.66 Å backbone RMSD

FIGURE 3.3: Best pairwise backbone alignments of structures calculated with Xplor-NIH (red) and DISCO (green).

for residues 29-37 likely due to increased dynamics in this region. The observed chemical shifts of the foldon domain in gp41-M-MAT were consistent with those reported for trimerized foldon, confirming that the foldon was folded and trimerized (Güthe et al., 2004). The gp41-M-MAT on the micelle exhibits unique architecture when compared to other gp41 trimer designs that incorporate non-native C-terminal trimerization domains (Frey et al., 2008; Hinz et al., 2009; Lenz et al., 2005; Liu et al., 2009).

The 2F5 epitope core residues D40, K41, and W42 are highlighted in Figure 3.2 B and C showing they are accessible for antibody binding. Residues D40 and K41 are on the surface of gp41-M-MAT, with part of the W42 side chain oriented towards the axis of symmetry leaving the rest at least partially solvent exposed. The

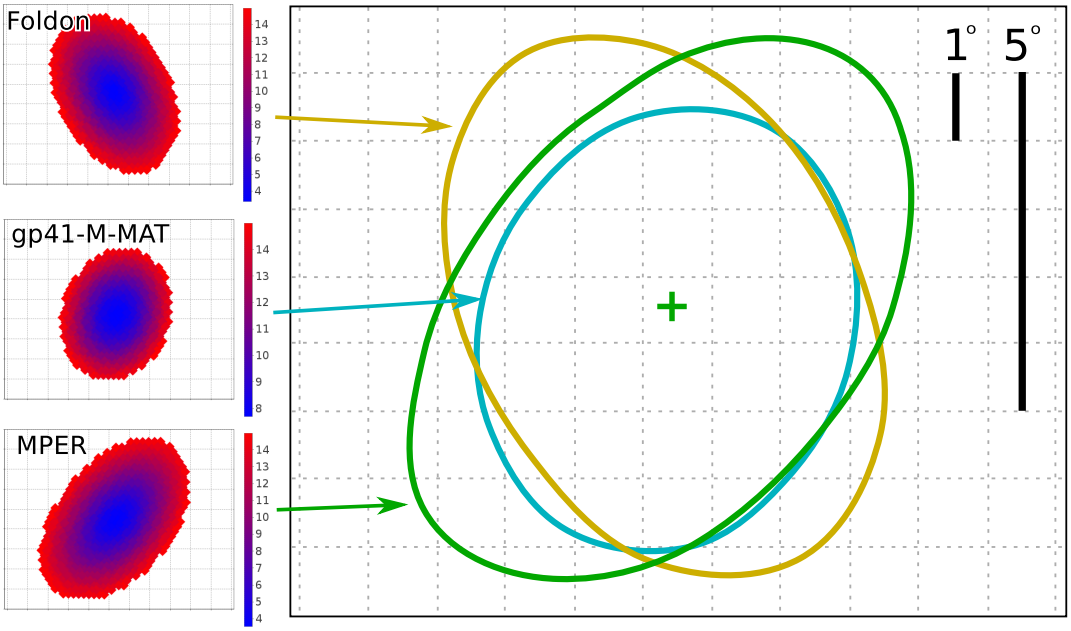
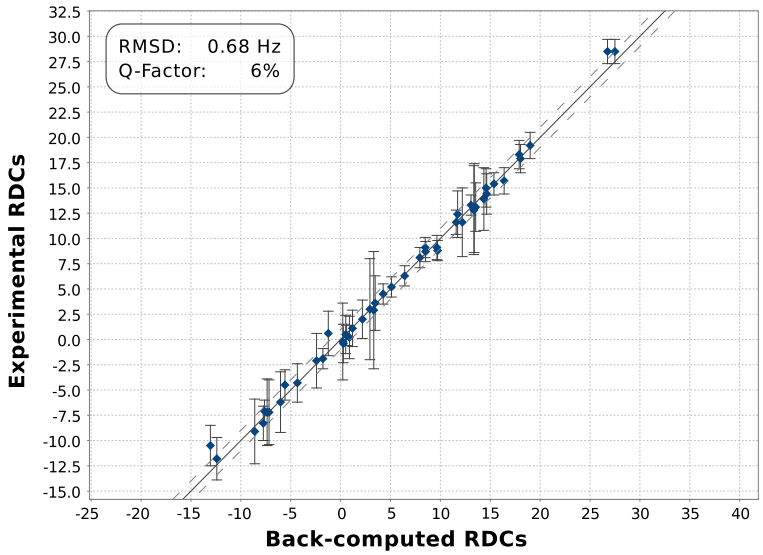


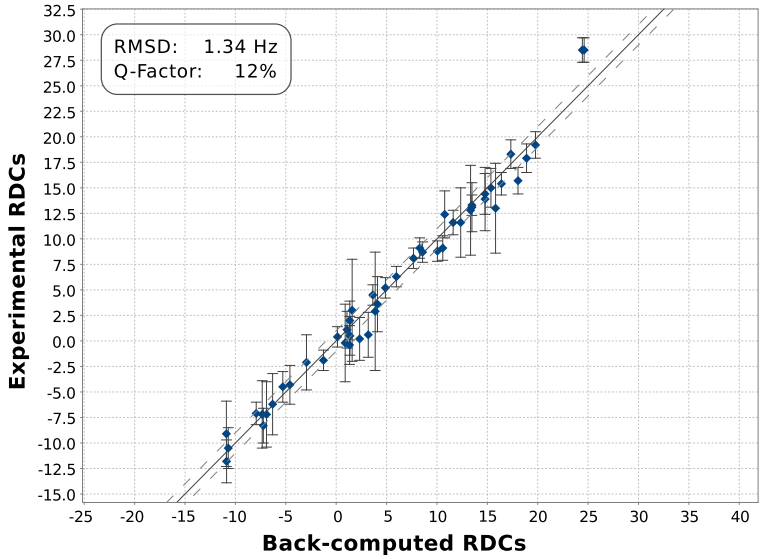
FIGURE 3.4: Symmetry axis orientations computed from the RDCs and the MPER subunit structure are the same as the ones computed for the full-length gp41-M-MAT and Foldon subunit structures. Left: Symmetry axis orientations sampled from the sphere are scored by how well the corresponding alignment tensor fits the structure. Scores are shown in color using RDC Q-factor. Only orientations whose RDC Q-factor is less than 15% are shown, which is only a small section of the sphere. Samples were projected to a tangent plane for display, where each grid cell is 1 degree wide. Since the sizes of these sections are small, the distortion due to the projection is also small. Right: Overlay of the boundaries of the symmetry axis orientations with RDC Q-factor less than 15% for the three structures. The green plus marks the best symmetry axis orientation for the MPER and the one used in DISCO's calculations.

individual members of the calculated ensembles from both 2LP7 and 2M7W exhibit varying degrees of solvent exposure for W42 due to the limited side chain constraints observed. Interestingly, the least conserved residue in the 2F5 core epitope, K41, is oriented directly away from the trimer interface towards the bulk solvent. The structural conformation of the 2F5 epitope in gp41-M-MAT exhibits little similarity with that observed in crystal structures of short MPER peptide segments bound to 2F5 Fabs (Ofek et al., 2004).

The 4E10 epitope region, W48 – T52, highlighted in Figure 3.2 B and D, is helical in our NMR structure. In the crystal structure of 4E10 bound to short



(a) Xplor-NIH gp41-M-MAT subunit structure



(b) DISCO-based gp41-M-MAT subunit structure

FIGURE 3.5: Experimental vs Back-calculated RDCs for the two GP41-M-MAT subunit structures. Due to symmetry, the fit of the RDCs to the subunit is the same as to the trimer structure. Both structures are in excellent agreement with the RDCs. Error bars on the points indicate the error of the experimental RDC measurement. The diagonal line indicates the region of perfectly-matching RDC values. The dotted lines along the diagonal indicate a distance of 1 Hz from the diagonal.

epitope containing peptides, W48, F49, I51, and T52 are in a helical conformation and account for the most contacts between the peptide and the antibody (Cardoso et al., 2005). Figure 3.2 D shows the side chain of F49 in our construct is directed inward toward the axis of symmetry, potentially contacting the micelle (discussed below). This is similar to the orientation proposed for F49 in the linear MPER peptide monomer structure, 2PV6, which suggested that F49 is buried in the lipid or micelle (Sun et al., 2008). N50 is exposed on the outer surface and is somewhat less conserved than the other residues in the epitope (Cardoso et al., 2005; Zwick et al., 2001). In the crystal structure, this residue makes fewer contacts with 4E10 when compared to most of the other residues in the peptide epitope (Cardoso et al., 2005).

3.2.2 Experimental characterization

Experiments performed by our collaborators, Patrick Reardon, Harvey Sage, and S. Moses Dennison probed the interaction of gp41-M-MAT with broadly neutralizing antibodies 2F5, 4E10, and 10E8. Analytical ultracentrifugation experiments revealed that 2F5 Fabs bound gp41-M-MAT presented on detergent micelles up to three times per trimer with a binding constant of ~143 nM per binding site. Surface plasmon resonance (SPR) experiments revealed that full-length antibodies bound tightly to gp41-M-MAT displayed on liposomes with binding constants of 0.18 nM for 2F5 and 27 nM for 4E10. These results indicate full-length 2F5 and 4E10 antibodies bound to gp41-M-MAT displayed on liposomes much more tightly than the Fabs bound to gp41-M-MAT displayed on detergent micelles. SPR data for full-length 10E8 indicated poor binding to gp41-M-MAT displayed on liposomes. We hypothesize that 10E8 targets the resting pre-fusion MPER state rather than the pre-fusion intermediate state of MPER captured by gp41-M-MAT.

NMR experiments performed by Patrick Reardon using a paramagnetic probe, 16-

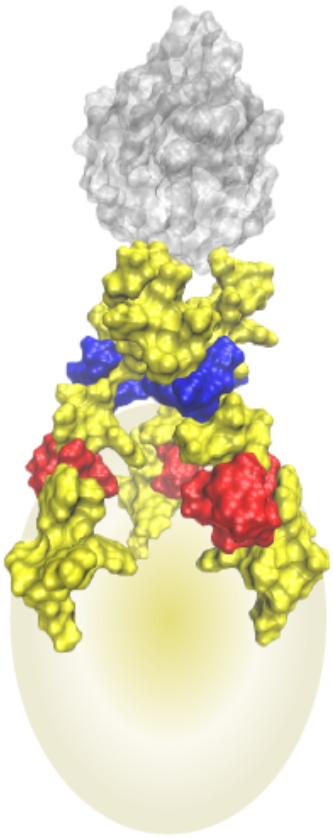


FIGURE 3.6: Schematic model for gp41-M-MAT association with a micelle. The micelle is represented by a prolate spheroid (yellow) containing the hydrophobic core with dimensions of 28 × 36 Å. This shape and size are based on previous small angle X-ray scattering measurements (Lipfert et al., 2007; Gobl et al., 2010) and is intermediate to the range of shapes reported.

DOXYL-stearic acid (DSA), found that residues 36 through 59 of gp41-M-MAT were in close proximity to the detergent micelle, indicating the C-terminal region of the MPER peptide in gp41-M-MAT associates with the detergent micelle, hypothetically by surrounding the micelle with the splayed conformation of the C-termini (See Figure 3.6). The AUC data indicate that gp41-M-MAT binds one micelle per trimer, which supports this hypothesis.

Heteronuclear NH NOE experiments, also performed by Patrick Reardon, showed the Foldon region of gp41-M-MAT is relatively rigid, while the MPER is relatively

dynamic. We concluded the N-terminal region of MPER, including the linker, experiences fast internal motion. The middle of the MPER is almost as rigid as the Foldon domain, but the backbone mobility increases at the MPER C-terminus.

3.3 Discussion

General physical properties and antibody binding characteristics of several gp41 MPER derived peptides have been reported, including at least three trimeric constructs in an extended conformation hypothesized to be similar to the prefusion intermediate state (Frey et al., 2008; Hinz et al., 2009; Lenz et al., 2005), the putative target of 2F5 and 4E10 antibodies (Frey et al., 2008). Two of these trimeric constructs are reported to induce antibodies that react with the MPER when administered to rodents (Hinz et al., 2009; Lenz et al., 2005); however, these antibodies are non-neutralizing.

Structural studies of the MPER domain have been limited primarily to short, monomeric peptide sequences and two trimeric constructs. The short peptides were either solubilized in detergent micelles (Sun et al., 2008; Schibli et al., 2001; Coutant et al., 2008) or in bound forms co-crystallized with MPER recognizing antibodies (Ofek et al., 2004; Cardoso et al., 2005). Of the trimeric structures, one was not membrane associated and does not bind 2F5 or 4E10 antibodies (Liu et al., 2009) and the other was part of a six helix bundle representing the post-fusion state of gp41 (Buzon et al., 2010).

The gp41-M-MAT construct characterized here does not contain the C-terminal transmembrane domain of gp41, but it does associate directly with micelles and liposomes. We note that in the virus, the trimerization state of the transmembrane domain that immediately follows the MPER domain is not well characterized. Some cryo-EM studies of intact Env show evidence for little or no trimerization of the the transmembrane domain (Zhu et al., 2006; Wu et al., 2010). Others are consistent

with a more compact structure having a width of ~35 Å where the Env stalk enters the membrane (Zanetti et al., 2006; White et al., 2010; Mao et al., 2012). The crystal structure of the post-fusion six-helix bundle shows that the MPER packs onto the outside of the bundle, leading to significant separation of the C-termini of the MPER domain (Buzon et al., 2010). In our structure, the C-termini of the MPER trimer segments are not self-associated, and instead are associated with the detergent micelle. Our structure may represent an intermediate state where the transmembrane domain is not tightly bundled allowing the MPER to associate with the viral membrane in conformations that enable its important function in the membrane fusion process (Muñoz-Barroso et al., 1999; Salzwedel et al., 1999). Thus, it will be of interest to determine the status of the MPER in atomic level structures of intact gp41-gp120 trimers.

Structure determination of multimeric membrane associated proteins in solution is challenging, and very few structures have been reported. In general the large size of the assembly and the unfavorable spin relaxation of systems like micelle solubilized proteins often limit the structural restraints observed in NMR. To determine the gp41-M-MAT structure, we combined traditional NMR structure determination techniques with novel methods based on residual dipolar couplings and intermolecular NOEs (Martin et al., 2011c).

Of the observed NOE restraints, only two of them were between subunits in the trimer. Two intermolecular NOEs would be insufficient to pack the helices for a general multimeric structure. Indeed, without any additional constraint, the space of possible packings between two helices has six degrees of freedom: rotations and translations, or $\mathrm{SO}(3) \times \mathbb{R}^3$ However, in the case of a symmetric homo-trimer, such as the MPER in gp41-M-MAT, after the orientation of the symmetry axis has been determined, the constraint imposed by the symmetry reduces the dimensionality to two degrees of freedom: positions of the symmetry axis relative to the subunit

structure, or $\mathbb{R}^2$ (Martin et al., 2011c). With two degrees of freedom instead of six, far fewer inter-subunit restraints are needed to define the helical packing. For the MPER trimer, we show in Figure 3.7 that two NOEs are sufficient to pack the interface.

The DISCO approach treats symmetry differently than Xplor-NIH. In Xplor-NIH, a potential that penalizes differences among the subunit structures, along with a potential encoding inter-subunit distance restraints, is used to implicitly represent the symmetry axis. In contrast, DISCO explicitly models the symmetry by computing the parameters of the symmetry axis directly, and hence defines a simpler structure determination problem. Under this parameterization, every possible quaternary structure satisfies the symmetry and all that remains is to compute the subset of symmetric quaternary structures that satisfy the experimental restraints. When the symmetry and the experimental restraints are simultaneously satisfiable, and when Xplor-NIH discovers satisfying quaternary structures without falling into a local energy minimum, the two approaches will return similar answers. The advantage of the parametric representation is that all satisfying quaternary structures can be reported, or it can be proven that none exist.

The limited number of side chain NOEs observed leads to some variation in side chain placement, especially for the more dynamic N-terminal region of the MPER. However, the Xplor-NIH and DISCO derived structural ensembles exhibit remarkably similar variations in side chain placement. This is important for the 2F5 epitope region where both ensembles contain structures that expose W42, and others where it is more buried in the interface with the trimer and possibly the micelle. These different conformations are all consistent with the NMR restraints, although the antibody binding data demonstrate that the W42 side chain is sufficiently exposed to support strong antibody binding.

Superposition of the gp41-M-MAT epitopes onto the respective 2F5 and 4E10

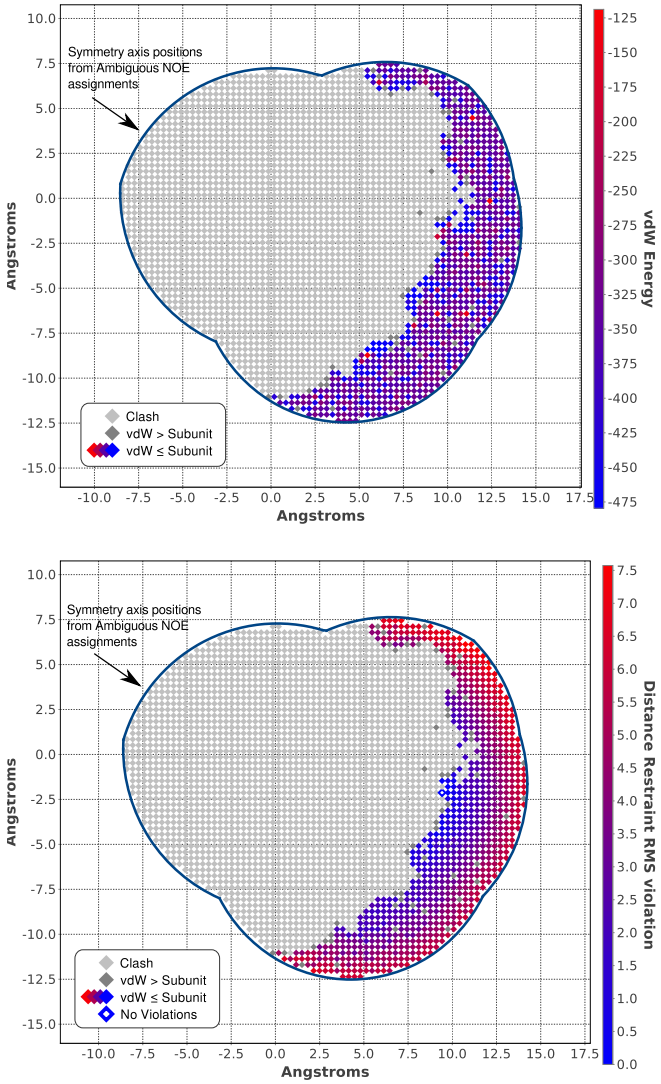


FIGURE 3.7: Analysis using symmetry can pack the MPER subunit:subunit interface with just two NOEs. Symmetry axis positions were sampled from the regions computed by DISCO using both the ambiguous (outer blue boundary) and the unambiguous (inner blue boundary) NOE assignments. Each axis position is represented relative to the subunit centroid (the origin) and is shown as a colored diamond. The axis positions are colored by (top) van der Waals energy and (bottom) NOE satisfaction. The axis positions with both favorable van der Waals packing (i.e., lower energy than MPER subunit alone) and low NOE violations occupy only a small region of space. The two plot axes are in angstroms and correspond to the plane shown in Figure 3.8(b). Different choices of the axis position correspond to different packings of the MPER trimer. Therefore, the score for each axis position is the score for the corresponding packed MPER trimer. The scores shown are from the single best conformer in the MPER subunit ensemble, but are representative of scores for all the conformers in the ensemble. Scores were computed after fixed-backbone local energy minimization using Xplor-NIH which could only vary side chain conformations.

Fabs at the epitope recognition sites in the crystal structures (Ofek et al., 2004; Cardoso et al., 2005), produced significant steric clashes with other parts of the MPER domain. In both 2F5 and 4E10, the residues N-terminal of their respective epitopes in gp41-M-MAT clash with large portions of the antibody. The clashes observed are extreme, and show that docking of the gp41-M-MAT structure onto the epitopes in the crystal structures does not produce a realistic representation of antibody binding to gp41-M-MAT. This suggests that the MPER domain in our construct undergoes significant conformational changes that turn the N-terminal region of the MPER away from the antibody upon binding. Importantly, gp41-M-MAT binds 2F5 and 4E10 with high affinity, demonstrating that the helical conformation in our NMR structure does not inhibit antibody binding. Instead, the increased dynamic flexibility observed in our structure at the 2F5 and 4E10 binding sites may allow the MPER sufficient mobility to alter its conformation upon antibody binding and turn the N-terminal region of the MPER away from the antibody to avoid steric clashes. This is important for HIV-1 vaccine development since a rigid vaccine candidate may not be able to mimic this behavior and consequently fail to induce 2F5 or 4E10 like antibodies. Furthermore, all three 2F5 epitopes on gp41-M-MAT trimer bind 2F5 with high affinity, making gp41-M-MAT a novel multi-valent antigen that effectively presents the MPER epitopes for recognition and high affinity binding.

3.4 Conclusion

The gp41-M-MAT structure is an antigenic, trimeric MPER domain directly associated with the lipid membrane without an exogenous trimerization domain at the C-terminus. It provides important structural information that can further illuminate HIV vaccine development efforts. Finally, the structure of gp41-M-MAT is an important addition to the relatively small number of multimeric membrane associated structures determined using solution state NMR.

3.5 Appendix: DISCO-based structure calculation

Using the NOE and RDC data collected from the trimeric gp41-M-MAT construct, structures of a single subunit of gp41-M-MAT were refined in isolation (i.e., without the other two subunits and including only intramolecular restraints) using Xplor-NIH (Schwieters et al., 2006). Of 100 conformers calculated, the 16 lowest energy structures with no NOE violations greater than 0.5 Å and no dihedral angle violations greater than 5° were chosen and the lowest energy subunit structure was selected as the best conformer.

We used the MPER fragment (residues 32–59) of this best conformer to compute the orientation of the three-fold symmetry axis of the trimeric structure that best fit the RDC data using the DISCO method (Martin et al., 2011c).

Briefly, DISCO computes the quaternary structure of homo-oligomeric proteins using assigned RDCs and distance restraints that can have ambiguous or unambiguous assignments. Suppose that coordinates for an ensemble of subunit structures has been computed using exclusively the intra-molecular restraints on the subunit structure. Then, taking this ensemble of subunit structures as input, DISCO is guaranteed to return all possible placements of the subunits (i.e., the packings) that are consistent with the symmetry, the RDCs, and the distance restraints. The packings are represented in terms of a parametric model of the symmetry axis of the oligomeric structure. Once the orientation and position of the symmetry axis relative to the subunit structure are computed, the full oligomeric structure can be reconstructed using the symmetry (Figure 3.8).

The alignment tensor resulting from DISCO's RDC analysis fit the RDC data and the MPER subunit structure with an RDC Q-Factor of 3.2%. The rhombicity of the alignment tensor was 0.02, which is consistent with the zero rhombicity expected for an axially-symmetric C_3 homo-trimer. The calculation did not use any coordinates

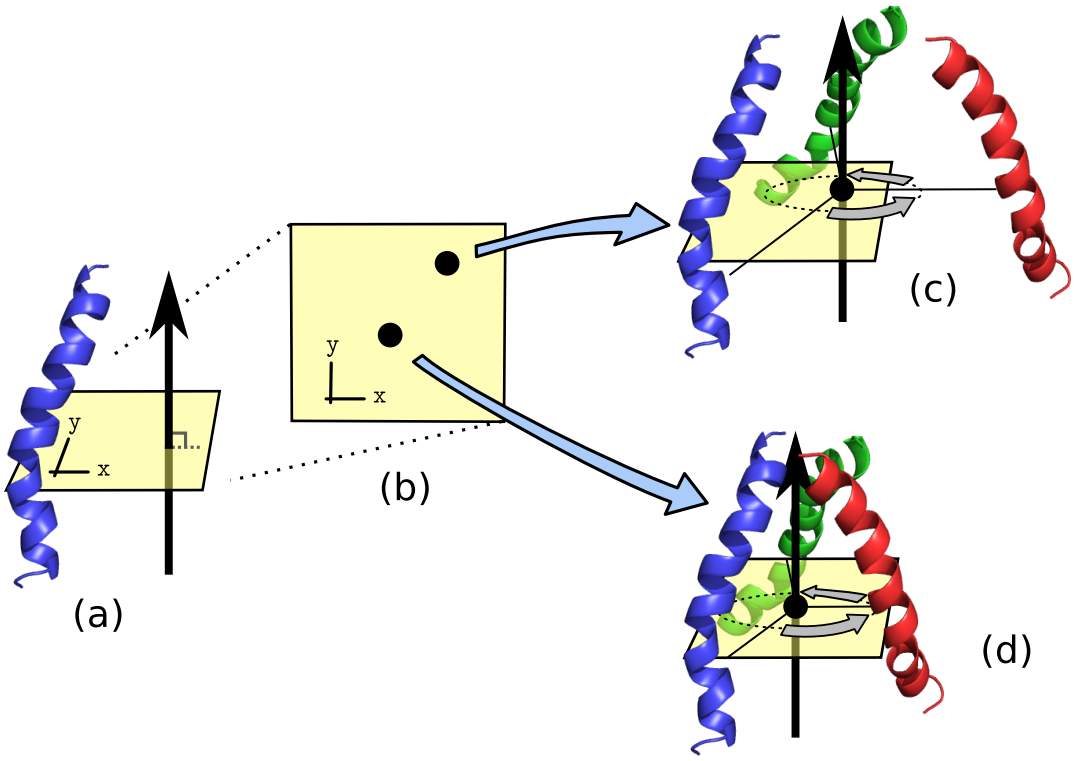


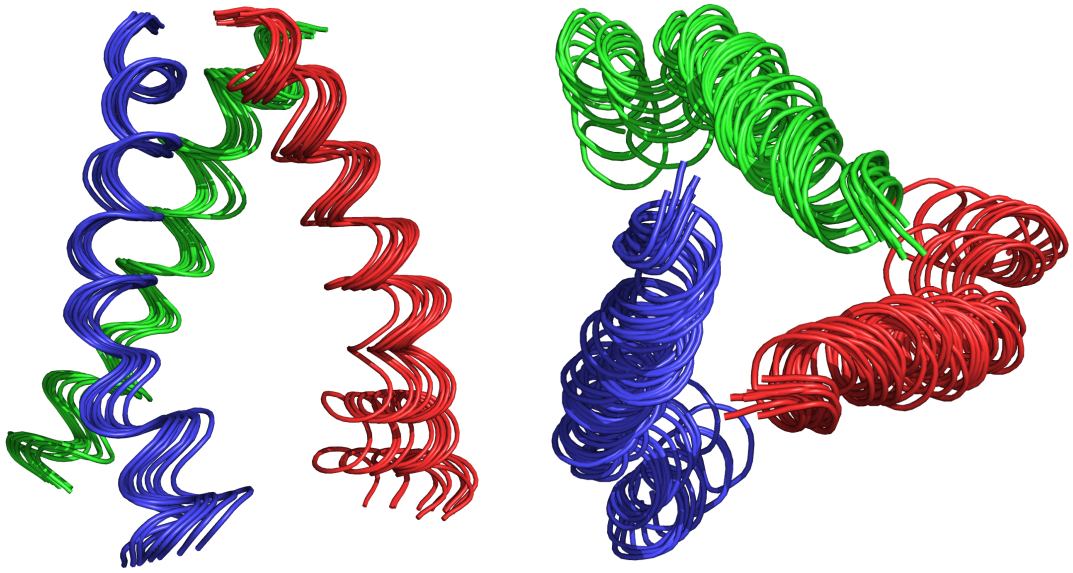
FIGURE 3.8: (a) The symmetry axis orientation (black arrow) is defined by the RDCs and defines a plane (yellow) with two degrees of freedom ($\mathbb{R}^2$) that encodes all the positions of the axis relative to the subunit structure (blue). (b) Plane of symmetry axis positions. Oligomeric structures are packed by applying the symmetry operations to the subunit structure. Different choices for the symmetry axis position define different oligomeric packings. (c) One hypothetical choice leads to a loose packing. (d) Another choice leads to a tight packing.

or RDCs from the Foldon fragment (residues 1–27) of gp41-M-MAT, and hence it is an independent calculation of the alignment tensor and symmetry axis orientation using the MPER subunit alone. In addition, a systematic search on a fine grid over all the possible symmetry axis orientations (Figure 3.4) did not reveal any satisfying orientations that were significantly different from the best-fit orientation, hence indicating the orientation of the symmetry axis is well-defined relative to the MPER subunit. The grid was constructed by subdividing the faces of a regular icosahedron seven times and projecting the resulting vertices onto the 2-sphere. Moreover, the orientation analysis yielded the same symmetry axis orientation for the Foldon frag-

ment of the best gp41-M-MAT subunit conformer as well as the full-length construct, hence showing the average orientations of the MPER and Foldon are axially aligned.

Using the computed symmetry axis orientation, the MPER fragments from the ensemble of gp41-M-MAT subunit structures, and the two inter-MPER NOEs, we computed the trimeric packing of the MPER using DISCO, which considered all possible packings allowed by all possible assignments of the inter-subunit NOEs due to chemical shift ambiguity and subunit assignment ambiguity. Although the chemical shift of the ILE 58 $H^{\gamma 2}$ proton was within the 0.05 ppm error window for one for the NOEs and hence was considered one of the assignment possibilities for that NOE, DISCO's analysis pruned the assignment using the MPER subunit structure, the RDCs, and the symmetry constraints. In other words, DISCO could prove this possible assignment was inconsistent with the RDC data and the gp41-M-MAT subunit structure, and therefore prune it from additional consideration. Using a van der Waals score and an NOE satisfaction score to measure the quality of possible packings, we found that two inter-subunit NOEs (resulting in six restraints on the trimer, due to symmetry) were sufficient to precisely define the trimeric packing (Figure 3.7). Figure 3.9 shows N-C^{α}-C backbone traces of the ensemble of MPER structures computed by DISCO. The 10 structures in the ensemble had zero violations of the two inter-subunit NOEs and had van der Waals packing scores (-470 to -118 kcal/mol) better than that of at least one of the MPER subunit conformations in isolation (-124 to -113 kcal/mol).

Since the MPER trimer computed by DISCO was calculated using only the MPER subunit structure and NMR data from the MPER, it represents an independent calculation (i.e., independent from the Xplor-NIH calculation) of the quaternary structure of the MPER. The Xplor-NIH and DISCO ensembles are essentially the same (Table 3.1) although the precise packing and side chain contacts in each case vary somewhat within the ensembles, likely due to the fact that the side chain conforma-



(a) side view (b) top view down the trimer axis

FIGURE 3.9: The ensemble of 10 MPER structures computed by DISCO using two inter-subunit NOEs.

tions are not completely specified by the NMR restraints. The best (lowest RMSD) pairwise alignment of full-length gp41-M-MAT conformers in the Xplor-NIH and DISCO ensembles is shown in Figure 3.3b.

Using the MPER trimer ensemble computed by DISCO, we used Xplor-NIH to determine the structure of the remaining linker (residues 28–31) and Foldon segments and thereby construct a full-length trimeric gp41-M-MAT structure. Xplor-NIH was configured to refine the full gp41-M-MAT in two steps.

The first step computed a crude approximation to the global fold by bootstrapping the linker/Foldon structure determination using a known structure of the Foldon region. Model 1 of the NMR ensemble 1RFO from the PDB was refined against the gp41-M-MAT restraints and was designated the Foldon reference structure. The structure determination of full-length gp41-M-MAT proceeded by starting with the following conformations. For each trimeric MPER conformation in the ensemble computed by DISCO, the linker and Foldon segments were attached to each MPER

subunit such that the linker was set to an extended conformation and the Foldon region was set to the Foldon reference structure. In this conformation, the three subunits of the Foldon region were separated in space due to the extended linkers. These starting structures were refined in Xplor-NIH using the standard annealing protocol with potentials for bond angles, improper angles, bond lengths, van der Waals, and favored/allowed Ramachandran regions to ensure proper geometry. Symmetry was enforced using a potential to minimize the RMSD between the subunits. All the experimental RDCs for gp41-M-MAT were used along with all the intra- and intersubunit NOEs. Additional potentials encoded restraints from $^{3}J_{HNHA}$ couplings and dihedral restraints from TALOS (Cornilescu et al., 1999). Finally, all intersubunit ^{1}H-^{1}H distances in the Foldon reference structure closer than 3 Å were collected into a simulated NOE potential which was used to drive the assembly of the Foldon region during the Xplor-NIH simulation. During the step 1 refinement, the MPERs were left completely fixed in space, the Foldon subunits were each treated as rigid but were still allowed mobility, and the linker regions were allowed complete flexibility. Each DISCO-based starting structure was refined 10 times and the lowest energy structure was accepted. The ensemble after step 1 had 10 members of full-length gp41-M-MAT with coarsely-defined linker and Foldon regions.

In the second step of refinement, all residues were set to flexible and mobile and the simulation was allowed to optimize for satisfaction of the experimental restraints to determine an ensemble of high-quality structures. The same potentials were used as step 1 with two exceptions. First, the potential encoding simulated NOEs for the Foldon region was replaced with an RMSD potential to the Foldon reference structure for the C^{α} atoms. Second, an RMSD potential to the original MPER trimer C^{α} atoms was added to minimize changes to the MPER, yet allow small changes to satisfy experimental restraints. Each of the 10 structures from step 1 was refined 20 times and combined to create a pool of 200 structures. Of this pool, 21

structures were selected that had no NOE violations greater than 5 Å, no dihedral angle violations greater than 5°, no scalar coupling violations greater than 2 Hz, and no NH RDC Q-Factors greater than 25%. Backbone clustering with a threshold of 0.5 Å RMSD was used to filter out duplicate structures which thinned the ensemble to 18 structures. These 18 structures were derived from eight of the original 10 DISCO-based MPER trimers, hence indicating the Xplor-NIH refinement and subsequent filtering pruned two of the trimeric MPER structures, presumably due to structural incompatibility with the trimeric linker and Foldon structures.

Essentially, the full gp41-M-MAT structure was computed by starting from the DISCO-based MPER timer, attaching the unfolded linker segments and the unassembled Foldon subunits, and then driving the Xplor-NIH simulation to recapitulate the Foldon trimer under the NMR restraints from gp41-M-MAT.

4

Systematic solution to homo-oligomeric structures determined by NMR

The text of this chapter has been adapted from a manuscript that was co-authored with Pei Zhou and Bruce R. Donald. The manuscript has not yet been published. In this section, my primary contribution is developing the fold-operator theory and applying it to solve many distinct structures of DAGK.

Abstract: Protein structure determination by NMR has predominantly relied on simulated annealing-based conformational search for a converged fold using primarily distance constraints derived from nuclear Overhauser effects (NOEs), PRE, and cysteine crosslinkings. Although there is no guarantee that the converged fold represents the global minimum of the conformational space, it is generally accepted that good convergence is synonymous to the global minimum. Here, we show such a criterion breaks down in the presence of large numbers of ambiguous constraints from NMR experiments on homo-oligomeric protein complexes. A systematic evaluation of the conformational solutions that satisfy the NMR constraints of a trimeric membrane protein, DAGK, reveals 10 distinct folds, including the reported NMR and

crystal structures. This result highlights the fundamental limitation of global fold determination for homo-oligomeric proteins using ambiguous distance constraints and provides a systematic solution for exhaustive enumeration of all satisfying solutions.

4.1 Introduction

Simulated annealing is a primary method for structure determination of proteins by nuclear magnetic resonance (NMR) spectroscopy (Schwieters et al., 2006; Herrmann et al., 2002). NMR restraints and biophysical principles are encoded into an energy function whose minimization results in models of the protein structure that satisfy the restraints. If the method consistently returns similar structures that adequately satisfy the restraints, the structural ensemble is considered well-converged and the structure determination is deemed successful. The main strength of simulated annealing is its ability to transform a coarse structural model into a more refined structure with improved restraint satisfaction. Where the method falls short is its ability to exhaustively sample topologically distinct structural models. Therefore, it can become trapped in the local minima of the energy landscape, thus missing the genuine fold(s) with similar or lower energies. Further complicating the situation, even if the global minimum structure of the energy function could be obtained, small inaccuracies in the energy function (e.g. due to approximation of complex physical phenomena or misinterpretation of even a few experimental distance constraints) could cause a genuine fold to be incorrectly ranked with a higher energy than the erroneous folds. Although such a situation is considered rare when all distance constraints are uniquely assigned, the odds increase significantly in the presence of ambiguous distance restraints for structure determination of homo-oligomeric protein complexes.

Ambiguous distance restraints (ADRs) (Nilges et al., 2010) refer to distance information (such as NOEs) that cannot be uniquely attributed to a single pair of

atoms. Since the chemical shifts of equivalent atoms in all subunits in a homo-oligomeric complex are identical and thus indistinguishable, ADRs are unavoidable for distance measurements in trimers and high-order homo-oligomers. We refer to this phenomenon as *subunit ambiguity* (Potluri et al., 2006, 2007; Martin et al., 2011c; Donald, 2011). Although it has been demonstrated that genuine interactions can be extracted from ADRs using an average distance function derived from a mean field approximation that encompasses the contribution of all degenerate atom pairs, such a method relies heavily on the initial fold for removing assignment ambiguity, which itself falls victim to the energy landscape of homo-oligomers containing a large number of minima with similarly low energy.

This situation is further exacerbated in the case of homo-oligomeric membrane proteins, for which dense restraint collection is often impractical (Vinogradova et al., 1998; Gautier, 2013; Bellot et al., 2013; Arora, 2013; Donald, 2011). In the case of Diacylglycerol Kinase from *Escherichia coli* (henceforth, simply DAGK), a membrane-associated homo-trimer, two different structures have been published. The solution NMR structure (Van Horn et al., 2009) of DAGK possesses a domain-swapped subunit interface, while the crystal structure (Li et al., 2013) has a subunit with a more compact conformation and without domain-swapping.

Here we show that the difference between the two structures is due to the local minimum limitations of current methodology for NMR structure determination. We demonstrate that this limitation can be mitigated by searching over topologically distinct folds using a systematic approach called *fold-operator theory*. Once an initial satisfying fold is discovered, mathematical operators transform the fold into alternate folds. The operators define a group action on the configuration space of protein folds. These alternative folds can be subsequently refined using traditional simulated annealing methods and evaluated for restraint satisfaction. Using this systematic approach, we found 48 distinct folds of DAGK, among which 10,

including the published NMR and crystal folds, upon energy minimization, satisfied experimental restraints.

4.2 Results

4.2.1 Schematic representation of three-dimensional structure exposes helical packing

To clearly show the differences in helical packing between the NMR and crystal structures (PDB IDs, respectively: 2KDC, 3ZE4), we reduced the three-dimensional structures of DAGK to two-dimensional *fold schematics* (Figure 4.1). From these schematic representations of the folds, it is easy to visualize the domain-swapped configuration of the NMR structure relative to the compact subunits of the crystal structure.

Of the deposited restraints collected for DAGK in solution, there are no inter-subunit NOEs, nor long range ($i-j>4$) NOEs within the same subunit. Hence, the NOEs, hydrogen bond restraints, dihedral angle restraints, and RDCs primarily constrain secondary structures within each subunit. The helices SH, H1, H2, and H3 are well-restrained individually, but the inter-helical linkers are relatively unrestrained, with little long-range information to pack the quaternary structure. The helical packing of DAGK, and hence the overall fold, is largely defined by the inter-subunit restraints: cysteine cross-linking via disulfide bonds, and restraints from paramagnetic relaxation enhancement (PRE). Since our previous analysis of the PREs showed that these long range constraints did not provide sufficient information to define the helical packing for DAGK (Martin et al., 2011c), we focused on the effect of cysteine crosslinking constraints and only used the PRE restraints as a filter to eliminate the erroneous structures.

The 24 disulfide bonds per subunit each have two possible subunit assignments for a homotrimer, and therefore the total number of assignment possibilities is 2^{24}, or

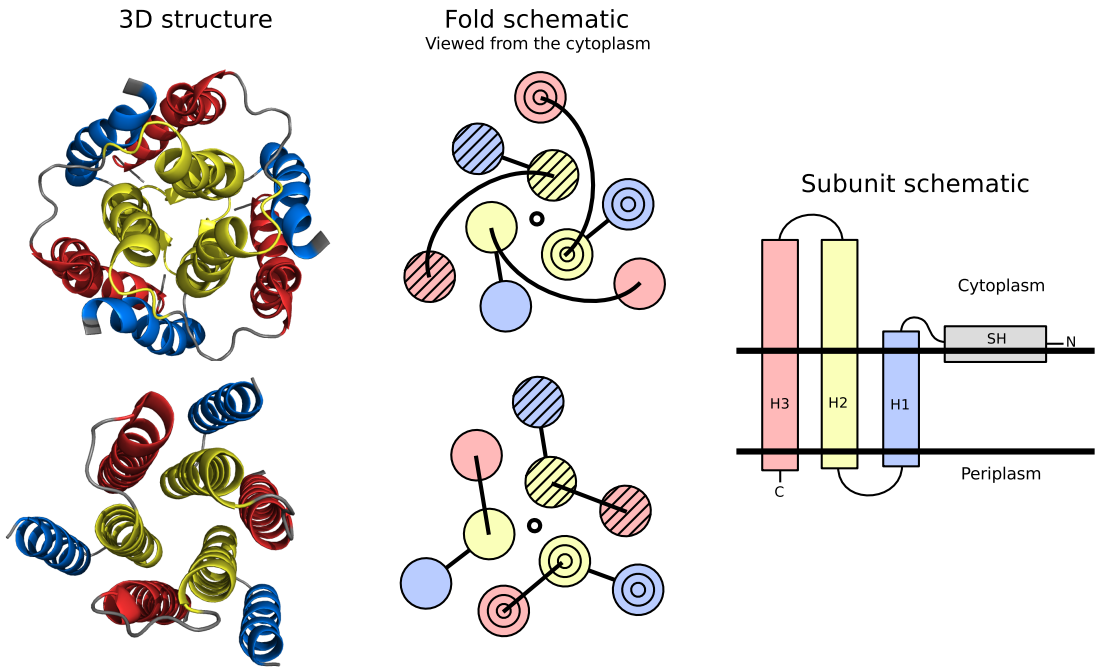


FIGURE 4.1: Fold schematics clearly show helical packing for the NMR (top) and crystal (bottom) structures of DAGK. In the fold schematic, the helices are shown as colored discs (the amphiphilic surface helix SH is not shown), the loop regions are shown as black lines, and the position of the three-fold symmetry axis is shown as a small black circle. Individual subunits are distinguished with different shading. Right: schematic of the subunit structure shows the helix naming and color schemes.

$\sim 17 \times 10^6$. However, the total number of assignment possibilities can be significantly reduced by classifying each of the 24 disulfide bonds into one of three categories depending on which pair of transmembrane helices was restrained: 8 in the H2-H2 category, 4 in H1-H3, and 12 in H2-H3. Since restraints in the H2-H2 category restrain the H2 helices to form a compact helical bundle in the core of DAGK regardless of the choice of subunit assignments, we focused on the contribution of the H1-H3 and the H2-H3 disulfide bond restraints. Restraints in each of the two remaining categories have two possible assignments each. Serendipitously, the two published structures satisfy opposite assignments: one assignment is satisfied by the NMR sturcture (hence referred to as the 2KDC assignment) and the other assignment is satisfied by the crystal structure (hence the 3ZE4 assignment).

When the assignments of all restraints in the H1-H3 and H2-H3 categories are synchronized, the total number of ways to assign the disulfide bonds drops from 2^{24} to just 4, and therefore it is feasible to examine each scenario individually. When all the disulfide bonds are set to the 2KDC assignments, we refer to this as the 2KDC *assignment scenario*. Alternatively, setting all disulfide bonds to the 3ZE4 assignments results in the 3ZE4 assignment scenario. Setting the H1-H3 category to the 2KDC assignments and the H2-H3 category to the 3ZE4 assignments results in the altA scenario, and the opposite assignments result in the altB scenario. The altA and altB assignment scenarios encode unreported additional structural solutions.

4.2.2 Fold-operator theory finds alternative folds allowed by restraints

Since the restraint provided by the disulfide bonds is ambiguous and rather loose ($\mathrm{d}_{\mathrm{C}_\alpha}(i,j) \leq 10$ Å), there are ways that the fold of the NMR and crystal structures for DAGK can be significantly changed without violating any disulfide bond restraints. For example, Figure 4.2 shows a sequence of changes that transform the crystal fold into the NMR fold, where the start fold, the end fold, and the intermediate fold all satisfy at least one assignment of each disulfide bond restraint.

The two changes described in Figure 4.2 can be decomposed into sequences of smaller changes called *operators*. These operators describe small changes to the folds that always result in a three-helical H2 bundle in the core of DAGK, and a maximal number of pairs of adjacent helices (i.e., the helical packing produced doesn't have holes in it), but don't necessarily produce only folds that satisfy the disulfide bond restraints. These operators are a mechanism to search the space of possible helical packings for DAGK to produce a set of folds which can be subsequently filtered against the disulfide bond restraints to return satisfying structures.

Only two operators, *roll* and *swap*, are needed to describe all the changes that can be made to the folds (Figure 4.3), and the application of all possible sequences

of these operators to the original NMR fold results in 48 unique possible folds for DAGK (Figure 4.4). For the example shown in Figure 4.2, the first change is the roll operator applied twice. The second second change is the swap operator applied once. Therefore, to transform the fold of the crystal structure into the NMR fold, one needs to apply the operator sequence `RRS` to the crystal fold where `R` is the roll operator, and `S` is the swap operator. These operators can be applied in any order and the result is the same. Consequently, `R` and `S` form the basis of a finite Abelian group of order 36.

4.2.3 Mathematical structure of the operators

The folds and operators for DAGK have a mathematical structure that provides a systematic way to sample topologically distinct folds and also precisely models the symmetry inherent to many homo-oligomers. The roll and swap operators (`R` and `S` respectively) form the basis set of an Abelian group G where the group operation is binary composition of operators. G has order 36 and the presentation

$$\langle \mathtt{R}, \mathtt{S} | 6\mathtt{R} = 0, 6\mathtt{S} = 0 \rangle \,, \tag{4.1}$$

so the canonical form of G due to the structure theorem for finitely presented Abelian groups Hungerford (1980); Donald (2011) is $\mathbb{Z}_6 \oplus \mathbb{Z}_6$. Here, $\mathbb{Z}_p$ denotes the ring of integers modulo p, also written $\mathbb{Z}/p\mathbb{Z}$. There are 36 operators in the group, yet there are 48 distinct folds for DAGK due to the multiple possibilities for linker routes after the helices have been placed.

Interestingly, `R` and `S` each generate a cyclic sub-group of order 6 (i.e., $\mathbb{Z}_6$) which is decomposed into $\mathbb{Z}_3 \oplus \mathbb{Z}_2$. The order-3 torsion subgroup reflects the trimeric quaternary structure of DAGK, while the order-2 torsion subgroup reflects the two remaining positions (modulo symmetry) for the H3 helix after the H1 helix has been placed (i.e., $3 - 1 = 2$). Therefore, the factorization of G into torsion subgroups of

prime order is

$$G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_2. \tag{4.2}$$

The group action of G on F, the set of folds, is the function

$$G \times F \rightarrow F \tag{4.3}$$

$$(g, f) \mapsto g \cdot f \tag{4.4}$$

where $g \in G$, $f \in F$, and $g \cdot f$ denotes operator g applied to fold f.

4.2.4 Predicted folds refine to satisfying structures

Starting from the 26 folds predicted by the fold-operator theory that satisfy the disulfide bond restraints, a starting structural model was built for each fold by using the fold as a structural template. These starting models were used as "seed" structures for later refinements using Xplor-NIH (Schwieters et al., 2003) against all available experimental restraints, including the PREs. The refinements were repeated 64 times for each fold to generate structural ensembles. Since some of the folds were predicted to satisfy two assignments of the disulfide bond restraints, the corresponding seed structures were refined twice – i.e., once for each assignment. Therefore, the 26 folds for DAGK resulted in 34 different ensembles. To simplify comparisons between the 34 different ensembles, we only report statistics on the lowest energy structure from each ensemble. As another simplification to aid comparison of many ensembles and also to report on the performance of simulated annealing, we report statistics on the convergence of each ensemble instead of using quality filtering that would normally be part of a standard structure determination protocol. The 34 resulting structures are shown in Figure 4.5.

Structures were evaluated using four measures. The first entails the *Xplor total energy* as the value of the energy function returned by Xplor-NIH after refinement

of individual structures, including the published NMR structure. Since all structures were refined using the same script, Xplor total energies are comparable across different structures (Figure 4.5).

The second scoring measure, the *RMS violation index*, is an RMS function of individual violation indices. Each violation index quantifies the satisfaction of a structure with respect to a class of restraints: NOEs, hydrogen bond restraints, RDCs, dihedral angle restraints, disulfide bond restraints, and PREs. Each violation index reports only the magnitude of the worst violation among the restraints in the class and is normalized from zero to one, where zero indicates perfect satisfaction of the restraints and one indicates the worst violation is equal to the chosen normalization constant. The normalization constants chosen for the violation indices in this study were: 0.5 Å for NOEs, 0.5 Å for hydrogen bonds, 1.0 Hz for RDCs, 5° for dihedral angle restraints, 2.0 Å for disulfide bond restraints, and 2.0 Å for PREs. Therefore, an NOE violation index of one or less indicates the worst NOE violation is 0.5 Å or less. The normalization constants can thus be chosen intuitively and allow violation indices for different restraint classes to be combined via the RMS function into a single statistic that reports the overall restraint satisfaction for a structure. Figure 4.6 shows comparisons of the RMS violation index with the Xplor total energy for the 34 determined structures.

The third and fourth scores are, respectively, the width score and the convergence score. These measures characterize the reliability of the structure determination instead of how well the lowest-energy structure satisfies restraints. The *width score* is the distance between the two lowest-energy structures found after each refinement, where the distance is the backbone atom (N,C^{α},C') RMSD in Å computed for helices H1, H2, and H3 (residues 30-48, 51-83, and 90-119) only. The width score reports when the structure determination returned a range of low-energy structures that are significantly different, or when the structures are all very similar. The *convergence*

score is the difference in Xplor total energy for the two lowest-energy structures found after each refinement. A low (i.e., good) convergence score means the lowest-energy structure is not merely an outlier or a "lucky guess" and that the structure determination can reliably return structures of similarly low energy.

In some cases, structures designed from one fold changed to another fold during refinement since we configured Xplor-NIH to perform full simulated annealing instead of just local energy minimization. There are 12 such switches in total, which are shown with brown arrows in Figure 4.5. When viewed as a dynamical system, the network of fold switches has four distinct attractors at folds M (the crystal fold), O (the NMR fold), B, and E (blue letters in Figure 4.5); folds B and E are not related to any published structures. The best seven structures by Xplor total energy and the best eight structures by RMS violation index were either seeded from, or switched to, one of these four attractor folds.

In the eight cases where a fold was predicted to satisfy two different assignment scenarios, the fold was refined twice, resulting in a pair of structures. One might expect each of these structures to resemble its partner, yet the distances between structures in the pairs (shown in red in Figure 4.5) are remarkably large. Structures in six of the eight pairs (C.altB, D.3ZE4, K.3ZE4, P.2KDC, Q.altA, and X.altA) changed to a different fold during refinement, so it is no longer reasonable to expect these structures to be similar to their partners. For the remaining two structure pairs (J.altB/J.3ZE4 and W.2KDC/W.altA), one of the two structures in the pair has a poor convergence score. Therefore, the distance between the two structures in the pair is not large compared to the width score.

The fold-operator theory for DAGK assumes the linkers connecting helices are arbitrarily flexible and therefore are able connect helices in any situation – even around other intervening helices. In reality, the linkers may not be that flexible and such strained folds are not kinematically feasible. Among the 12 cases where the

refinement allowed strained structures to escape to a more favorable fold, P.2KDC escaped to fold O, but its partner structure, P.altA, was "locked in" to the strained linker conformations of fold P by the disulfide bond assignments. Consequently, P.altA has the third-worst Xplor total energy and the second-worst violation index of all the structures.

Of the 26 folds predicted by the fold-operator theory for DAGK to be satisfying, 10 of these folds yielded at least one structure that met the expectations (on average) for restraint satisfaction by having an RMS violation index of 1 or lower. 6 of the 26 folds yielded structures that switched to different folds during refinement, so it is not known from these results if these 6 folds describe satisfying structures or not. 10 folds resulted in structures with RMS violation indices of greater than 1, and hence these structures did not meet expectations for restraint satisfaction. Figure 4.7 shows all the structures grouped by their post-refinement fold, and a full listing of the violation indices for each structure is given in Table 4.1.

4.3 Discussion

In many respects, the 2D schematic representation used in the fold-operator theory for DAGK is an oversimplification. Condensing the full three-dimensional structure of DAGK into a flat projection ignores some important structural details of DAGK. For instance, the transmembrane helices need not be strictly parallel, or even straight. Modeling changes to helix shape with operators could potentially enable the discovery of more satisfying folds, but simulated annealing methods likely already adequately search over such changes in helix shape. Since simulated annealing is prone to becoming stuck in local minima (like all local minimization methods) and therefore might miss genuine solutions, the goal is to choose operators that complement simulated annealing and overcome its local minimum limitations rather than use operators to model small changes to helix shape. Indeed, despite the sim-

Table 4.1: Violation indices for all refined DAGK structures

Structure name	NOEs 0.5 Å	HBonds 0.5 Å	RDCs 1.0 Hz	Dihedrals 5°	DBonds 2.0 Å	PREs 2.0 Å
2KDC[1]	0.35	0.12	0.7	0.45	0.24	0.23
A.altB	0.4	0.13	1.08	0.36	0.24	0.54
B.altB	0.35	0.14	0.43	0.34	0.23	0.63
C.3ZE4	0.56	0.46	2.99	0.87	0.45	0.47
C.altB	0.31	0.16	0.36	0.3	0.23	0.53
D.3ZE4	0.61	0.26	1.84	1.11	0.32	0.67
D.altB	0.53	0.39	2.59	0.86	0.24	0.65
E.3ZE4	0.47	0.2	0.94	0.76	0.3	0.41
F.3ZE4	0.51	0.29	2.13	0.77	0.26	0.62
G.altB	0.91	0.38	5.66	1.4	0.44	0.72
H.altB	1.21	0.28	2.99	1.19	0.47	0.46
I.altB	0.89	0.53	2.91	1.36	0.42	0.8
J.3ZE4	1.04	0.5	2.83	1.2	0.46	0.81
J.altB	1.22	0.31	2.25	1.02	0.23	0.64
K.3ZE4	0.62	0.31	2.61	1.42	0.26	0.67
K.altB	0.55	0.44	2.92	0.81	0.31	0.54
L.3ZE4	0.6	0.16	1.7	1.08	0.28	0.72
M.3ZE4	0.56	0.17	1.21	0.77	0.27	0.49
N.2KDC	0.49	0.2	1.04	0.54	0.27	0.73
O.2KDC	0.33	0.2	1.28	0.4	0.25	0.46
P.2KDC	0.38	0.13	0.32	0.51	0.29	0.5
P.altA	1.84	0.39	4.5	1.37	0.55	0.54
Q.2KDC	0.37	0.35	1.95	0.55	0.53	0.59
Q.altA	0.84	0.29	1.83	0.78	0.38	0.72
R.altA	0.85	0.31	2.85	0.69	0.36	0.73
S.altA	0.59	0.41	2.8	1.1	0.33	0.72
T.altA	0.64	0.27	1.33	0.61	0.4	0.52
U.2KDC	0.54	0.22	1.78	0.91	0.33	0.76
V.2KDC	0.39	0.28	2.22	1.11	0.26	0.59
W.2KDC	0.35	0.3	1.81	0.98	0.23	0.78
W.altA	0.51	0.34	4.34	1.24	0.31	0.68
X.2KDC	0.39	0.22	2.37	0.9	0.32	0.8
X.altA	0.54	0.2	2.16	1.45	0.34	0.78
Y.altA	0.59	0.34	2.1	1.05	0.32	0.61
Z.altA	0.62	0.29	2.11	0.89	0.26	0.44

[1]The published NMR structure, PDB ID: 2KDC, model 1. Violation index values are all unitless. A value of 1 or less indicates the structure meets expectations for restraint satisfaction. Normalization constants for the violation indices are shown under each column heading. Violation indices are further described in the main text.

ple representation of structure used by the fold schematics, the fold-operator theory predicted 24 distinct folds for DAGK that satisfied the disulfide bond restraints (in addition to the two published folds), of which 10 folds yielded structures that met stringent expectations for NMR restraint satisfaction.

The fold-operator theory presented here bears some similarity to methods in protein structure prediction. The *ideal forms* proposed by Taylor et al. (Taylor et al., 2008) describe different protein folds using the "combinatorial approach" (Cohen et al., 1980). Under this regime, possible folds are enumerated from a space of choices governing the placement of α-helices and β-sheets and then structures are fit to these ideal forms, refined, and finally scored. While our fold-operator theory shares the combinatorial generate-and-test approach, where the methods differ is how the combinatorial space is defined. The ideal forms were curated from a database of structural information, while in the fold-operator theory, the different folds are algebraically defined by the initial satisfying fold and the group action of operators.

We have demonstrated our method on DAGK, showing how to find a remarkable variety of satisfying folds, but the method can also be applied to other homo-oligomeric proteins where ambiguous restraints necessarily hinder structure determination with simulated annealing. The application of the fold-operator theory to a new protein requires defining F, a set of folds, and G, a group of operators, analogously to our example with DAGK. This defines a group action on the configuration space of folds (see SI). The first step is to discover one fold $f \in F$ that satisfies the restraints, and (similarly to our example in Figure 4.2) search the changes to the structure that preserve restraint satisfaction. If relatively rigid backbone fragments can be determined (e.g., helices within each subunit), then restraints can be categorized as restraining pairs of rigid fragments and the number total number of assignment possibilities is vastly reduced. Therefore, changes to f that preserve inter-subunit restraint satisfaction for symmetric homo-oligomers will generally include substituting

fragments in one subunit with identical fragments from other subunits.

The next step is to factor the satisfaction-preserving changes into a set of finer operators (e.g., Figure 4.3) that form the basis of an Abelian group G. The group structure is necessary to precisely model the symmetry inherent in many homo-oligomeric proteins, but the operators need not preserve restraint satisfaction. Removing this restriction was necessary to obtain the group structure in the case of DAGK, and, more generally, it allows the operators to hop between "islands" of satisfying folds. G and f are then used to construct F via the group action and therefore describe the possible folds. For DAGK, F was small and exhaustive search was a feasible method to find the low-energy folds. If F is large (which appears to require a larger protein than the $121 \times 3 = 363$ residue DAGK), more sophisticated algorithms may be needed, such as branch-and-bound pruning which is often used in protein design (Donald, 2011).

We have presented a general method for structure determination of protein homo-oligomers and demonstrated the method on DAGK. We conclude that the differences in the published NMR and crystal structures are due to limitations of current NMR structure determination methodology. When the convergence of a set of structures to a satisfying fold represents merely one of many possible folds allowed by ambiguous restraints, fold-operator theory allows systematic search over the space of possible folds. Using fold-space search methods to address the limitations of local minimization techniques such as simulated annealing enables robust structure determination for difficult homo-oligomeric systems, particularly membrane associated systems hindered by the availability of only sparse and ambiguous restraints.

4.4 Methods

4.4.1 Fold-to-structure protocol

To build atomic resolution structural models of DAGK, we first calculated a set of folds using the fold-operator theory. For each fold predicted to satisfy the disulfide bond restraints, we determined a structure of DAGK based on that fold using the following protocol.

1. Using PyMOL (Schrödinger, 2012), we created a reduced model of the DAGK subunit by deleting all but residues 6–12, 32–44, 50, 57–77, 85, and 94–117 from the PDB structure 2KDC, model 1. These residues are, respectively, fragments of the SH helix, the H1 helix, the H1-H2 linker, the H2 helix, the H2-H3 linker, and the H3 helix.

2. For a chosen fold, we translated and rotated the fragments from step 1 so they aligned with one subunit of the fold. This step created a template structure for the subunit of DAGK. Since the SH helix was not modeled by the fold schematics, the SH helix fragment was oriented so it pointed away from the core of DAGK.

3. Using Xplor-NIH (Schwieters et al., 2003), we annealed an extended (i.e., unfolded) model of a single DAGK subunit using the intra-subunit NMR restraints: NOEs, hydrogen bonds, dihedral restraints, and RDCs. We configured the refinement to penalize differences between the backbones of the refined model and the template structure created in step 2. The result was a structure of the DAGK subunit that simultaneously matched the chosen fold and satisfied the NMR restraints.

4. Using PyMOL again, we made three copes of the subunit structure created in step 3. We rotated and translated the subunit structures until they matched

the trimeric conformation of the chosen fold. The result here was a trimeric "seed" structure for DAGK to be used in later refinements.

5. Finally, we used Xplor-NIH to refine the trimeric "seed" structure from step 4 using all the experimental restraints: NOEs, hydrogen bonds, dihedral restraints, RDCs, disulfide bonds, and PREs. Unlike the subunit refinement, this trimeric refinement did not use a template structure to restrain the backbone of the refined structure. Without a backbone template, the trimeric refinement was free to change the fold of the structure when such a change resulted in a lower energy.

One drawback to the fold-to-structure protocol presented here is that unrestrained degrees of freedom are not necessarily sampled by the final ensemble. For instance, the SH helix in our ensembles appeared more converged than was suggested by the NMR restraints and as a result, the ensembles for the SH helix were falsely precise. Normally, unrestrained degrees of freedom are searched by the random structure generation used in the beginning of most annealing protocols. For small modes of variability, the random structural sampling is able to report a variety of structures, but has difficulty searching topologically distinct folds. The fold-operator theory presented here completely supplants random structural sampling as a mechanism to search alternate folds, so one must take care to ensure that all degrees of freedom are captured by the operators. In our case, variability in the SH helix had little impact on the fold of DAGK, so we chose not to model it using the operators.

4.4.2 Refinement using Xplor-NIH

In steps 3 and 5 of the fold-first structure determination procedure, Xplor-NIH v2.33 was used to refine structural models. The details of these refinements are described below.

We annealed subunit structures using 2000 steps of dynamics at 3000° K followed by a cooling phase where the temperature dropped to 25 K over 20,000 steps of additional dynamics. Then, models were minimized using 1000 steps of torsion angle minimization followed by 1000 steps of Cartesian minimization. Throughout the simulations, the models were restrained by the usual chemical potentials: bond, angl, impr, and the non-bonded atom repel potential. Also, the simulation used potentials for experimental intra-subunit restraints: NOEs, hydrogen bonds, dihedral restraints from TALOS, and RDCs. The potential used for the NOEs and the hydrogen bonds was the "hard" type NOE potential with the default averaging exponent of 6. The alignment tensor used for the RDC potential was restricted to have zero rhombicity, but its magnitude was allowed to vary. The orientation of the tensor was also fixed so its z axis aligned with the frame of the template structure, and hence the symmetry axis for the DAGK trimer. The C^{α} atoms of the subunit template structure were used to restrain corresponding C^{α} atoms in the models using an RMSD potential. Since these subunit structures were destined for further refinement in the trimeric state, only a single structure was calculated for each template. The single structures were found to have adequate satisfaction statistics and therefore computing large ensembles was deemed unnecessary.

For the subunit simulations, weights for the potentials were generally set to low values for the high-temp dynamics, and raised linearly during the cooling phase to final values, or simply held constant throughout the simulations. Table 4.2 shows the weights used for the subunit refinements.

Multiplicative ramps are often preferred in Xplor-NIH refinements. However, we observed during our refinements that potential weights were rising too quickly at low temperatures to allow the dynamics simulations to find relaxed conformations. Using a ramp with a slower rate of increase at low temperatures should have solved the problem, but the linear ramp gave poor results. Further investigation revealed

Table 4.2: Weights for the subunit refinements using Xplor-NIH.

Potential	High-temp weight	Low-temp weight
repel (radii scale)	0.4	0.8
repel (weight)	0.2	1
bond	1	1
angl	0.4	1
impr	0.1	1
NOE	100	100
hbond	100	100
dihedral	500	500
RDC	0.01	1
template RMSD	1	10

numerical round-off errors in the implementation of linear ramps. After implementing a numerically-stable linear ramp, cooling phases using the new linear ramps gave greatly improved results.

For trimeric refinements, we used a slightly different approach than the one used for subunit refinements. Instead of annealing an extended chain once, we refined the "seed" structures many times to compute a traditional NMR ensemble. We refined each of the 34 trimeric seed structures 64 times. From the resulting ensemble for a chosen fold, we chose the single lowest-energy structure to represent the fold. The convergence and width scores reported in Figure 4.5 were computed from this ensemble. We refined trimer structures using 4000 steps of dynamics at 3000° K followed by a cooling phase where the temperature dropped to 25 K over 40,000 steps of additional dynamics. Finally, models were minimized using 4000 steps of Cartesian minimization. The energy function was composed of the usual chemical potentials: bond, angl, impr, the non-bonded atom repel potential, and an additional RMSD potential between subunits to enforce the trimeric symmetry. The energy function also incorporated potentials for experimental restraints: NOEs, hydrogen bonds, dihedral restraints from TALOS, RDCs, disulfide bond restraints, and restraints from PRE. The potential for NOEs and hydrogen bonds was the "hard" type NOE poten-

Table 4.3: Weights for the trimer refinements using Xplor-NIH.

Potential	High-temp weight	Low-temp weight
repel (radii scale)	0.6	0.8
repel (weight)	1	1
bond	0.6	1
angl	0.6	1
impr	0.6	1
symmetry	5	5
NOE	20	40
hbond	20	80
dihedral	200	400
RDC	0.1	1
disulfide bond	1	40
PRE	0.01	30

tial with the default averaging exponent of 6. For the disulfide bond restraints and restraints from PRE, which are restraints with larger upper distances, we obtained better results by using an exponent of 12 with the "hard" type NOE potential. For the alignment tensor used in the RDC potential, the rhombicity was fixed at zero, but the magnitude and orientation were allowed to vary. Table 4.3 gives the weights for the potentials used during the trimeric refinement.

Empirically, we found that a weight on the order of hundreds was needed for the dihedral potential to have any noticeable effect. We also found that the weight for the PRE potential needed to be initialized with a very low value to prevent the simulation from becoming trapped in the many local energy minima defined by these very ambiguous restraints.

In a final step, the lowest-energy structure from each trimeric ensemble was subjected to an additional 4000 steps of Cartesian minimization. The conditions for this minimization were the same as for the previous low-temperature minimization with one notable exception. The van der Waals potential was used (with a weight of 1) instead of the non-bonded atom repel potential to drive molecular packing. We found that for structures already in a low-energy conformation, switching to the van

der Waals potential gave even better results. The values computed by the Xplor-NIH energy function after this final minimization are the energy values reported in Figure 4.5. The energy value reported in Figure 4.5 for the NMR structure of DAGK was computed using this final minimization as well to ensure that scores were comparable across different structures.

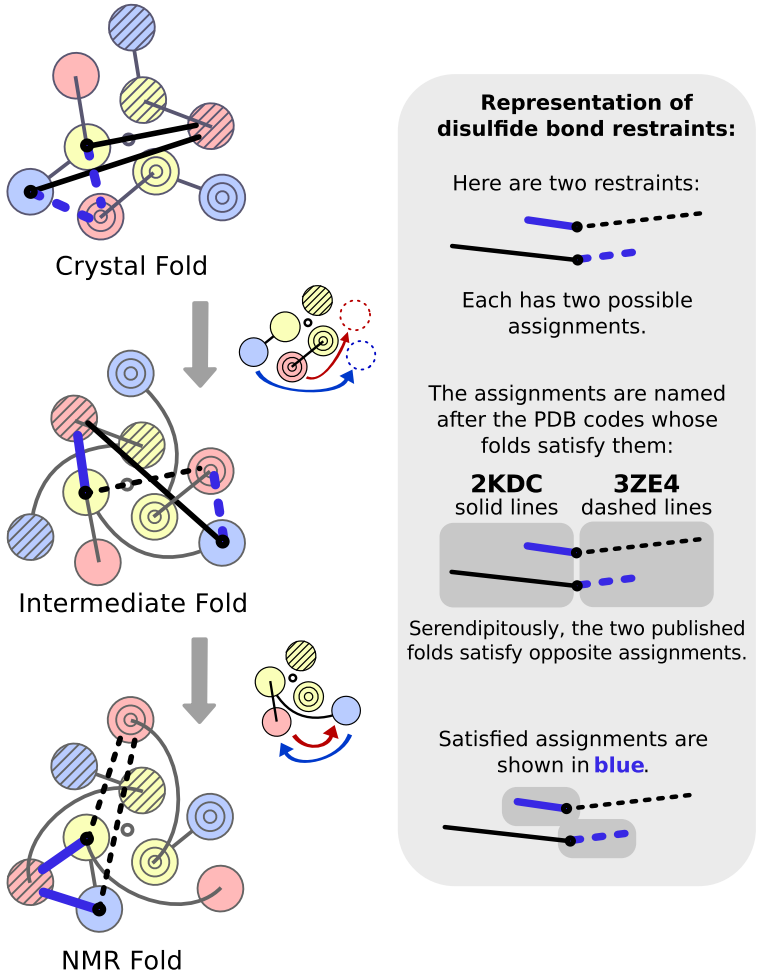


FIGURE 4.2: The crystal structure can be transformed into the NMR structure by repositioning the transmembrane helices. The changes are indicated by arrows. Top: In the fold of the crystal structure, the 3ZE4 assignments are satisfied, but the 2KDC assignments are not. Middle: Moving the H1 (red) and H3 (blue) helices as shown transforms the crystal fold into an intermediate fold that satisfies a mixture of 2KDC and 3ZE4 assignments, named the altB assignment scenario. Bottom: Swapping the H1 and H3 helices transforms the intermediate fold to satisfy the 2KDC assignments.

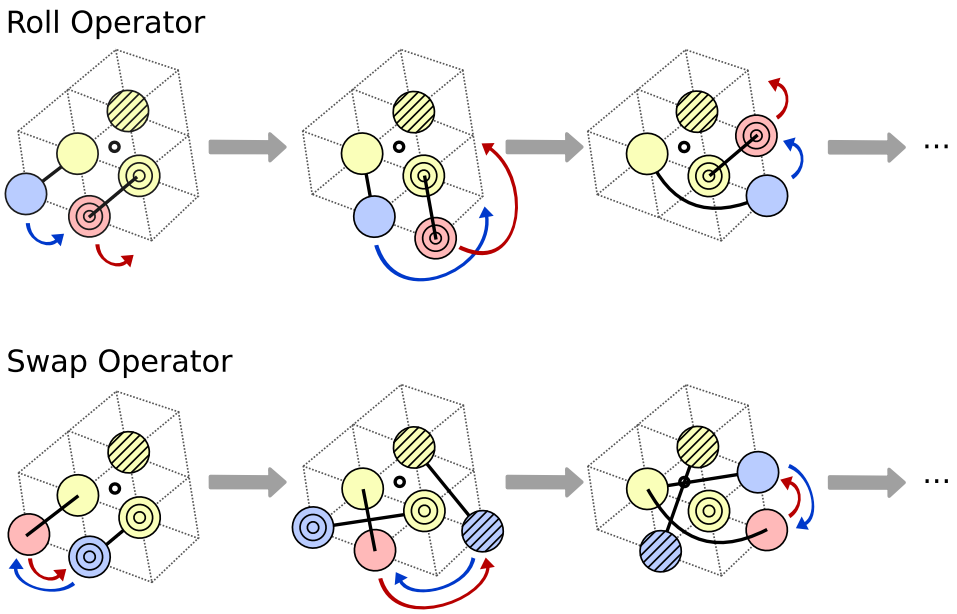


FIGURE 4.3: The two operators in the fold-operator theory for DAGK: The Roll operator moves the red and blue helices (H3 and H1 respectively) along the perimeter of the three-helix core (H2) in a counterclockwise direction. The Swap operator exchanges the position of the red helix (H3) with the blue helix (H1) that lies immediately counterclockwise adjacent to it. After six applications of either of the two operators, the ending fold is always the same as the starting fold.

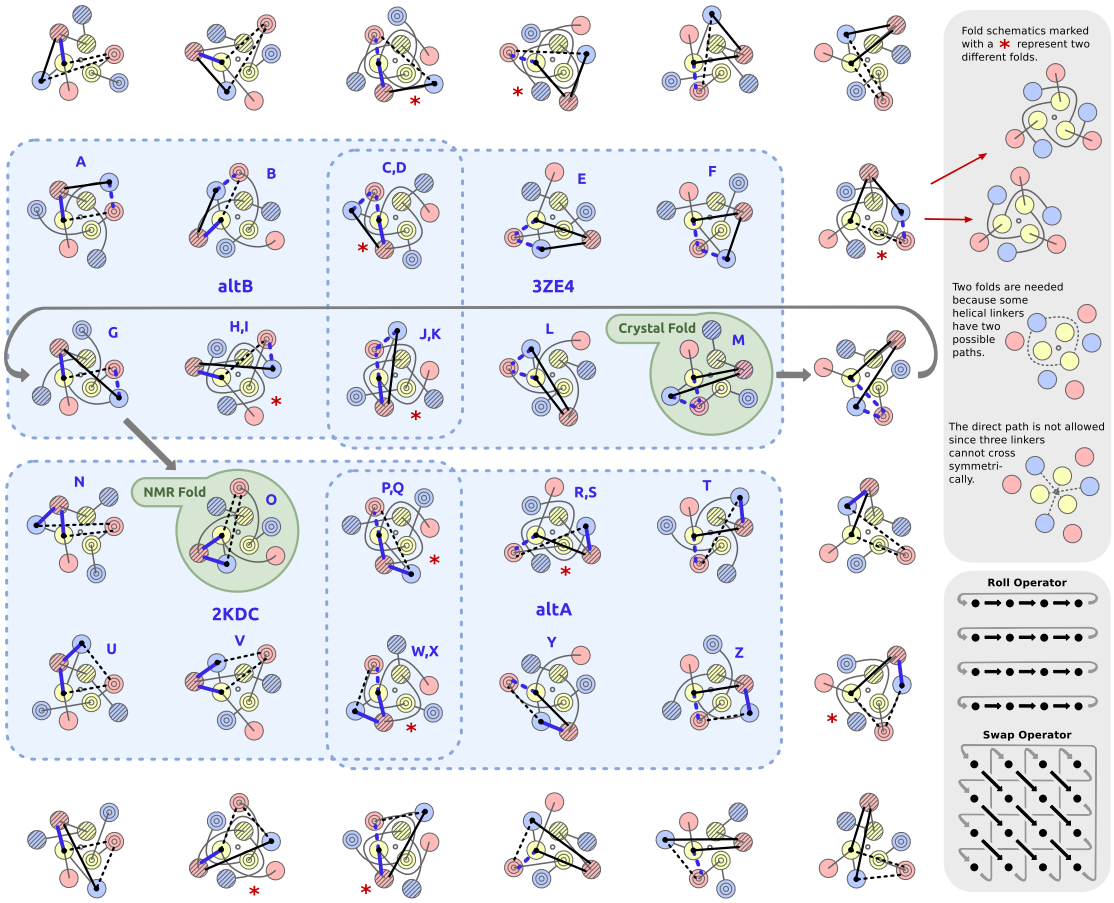


FIGURE 4.4: The fold graph of 48 distinct folds predicted for DAGK by the fold-operator theory. Graph vertices are represented by fold schematics. The edges are represented in the lower right panel. Generally, the roll operator sends any fold horizontally to its right neighbor. The swap operator sends any fold diagonally to its lower-right neighbor. Since the fold graph is embedded on the 2-torus, the operators "wrap around" the sides of the figure. Of these folds, 26 were predicted to satisfy the disulfide bonds, and 22 were not. The satisfying folds are grouped by the four assignment scenarios (2KDC, 3ZE4, altA, altB, shown with dashed blue boxes). Each satisfying fold was given a single-letter name, shown in blue. The operator sequence `RRS` that transforms the crystal fold into the NMR fold (also described in Figure 4.2) is shown with three grey arrows.

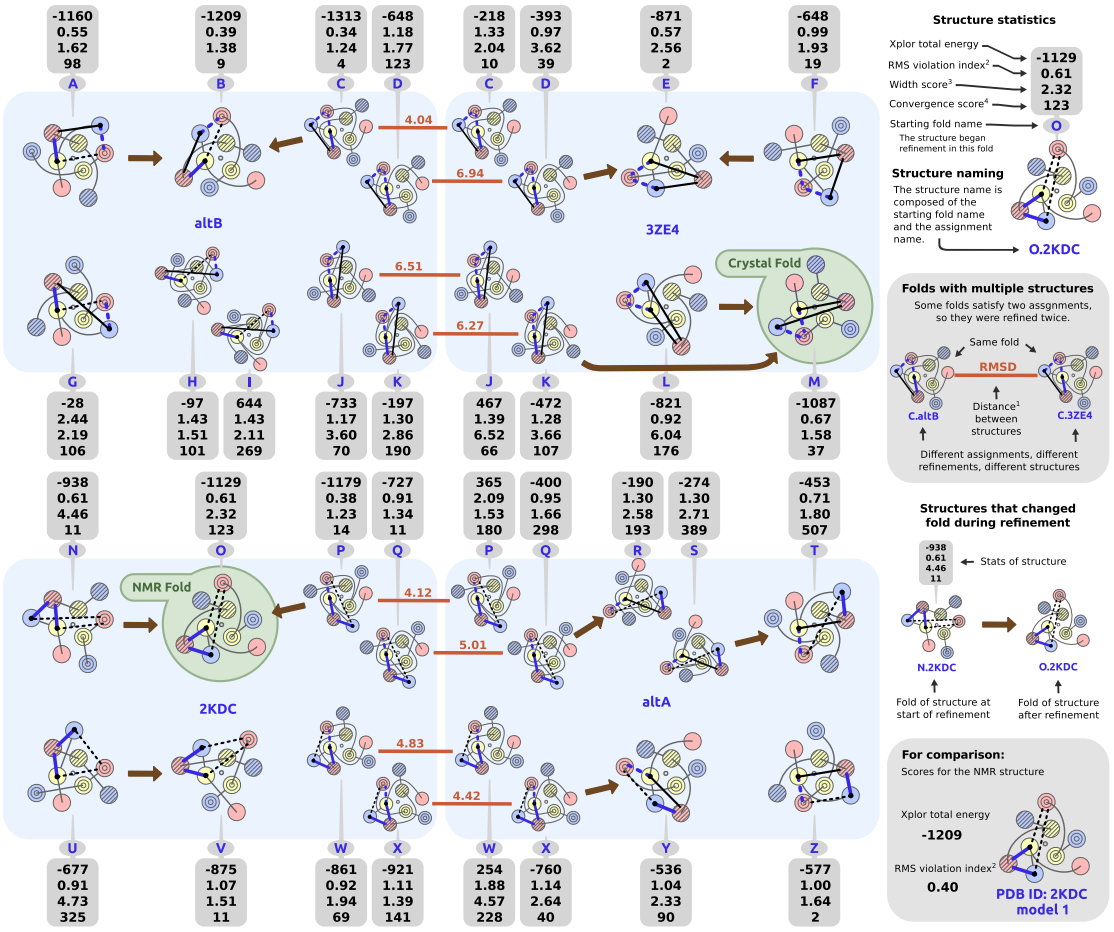


FIGURE 4.5: The 34 satisfying structures computed for DAGK. Each structure is shown using the schematic of the fold that was used to seed the refinement. Structures are grouped by the four disulfide bond restraint assignments (blue boxes). Structures that changed folds during the refinement are shown with brown arrows between the fold schematics. [1]All structural distances are backbone atom (N,C^{α},C′) RMSD values in Å computed for the helical residues 30-48, 51-83, and 90-119 only. Variations in the loop regions were not considered in this score. [2]The RMS violation index scores satisfaction of all solution restraints without regard to force field energies. This score is described in the text. [3]The width score is the distance[1] between the two lowest-energy structures computed for that fold. [4]The convergence score is Xplor total energy between the two lowest-energy structures computed for that fold.

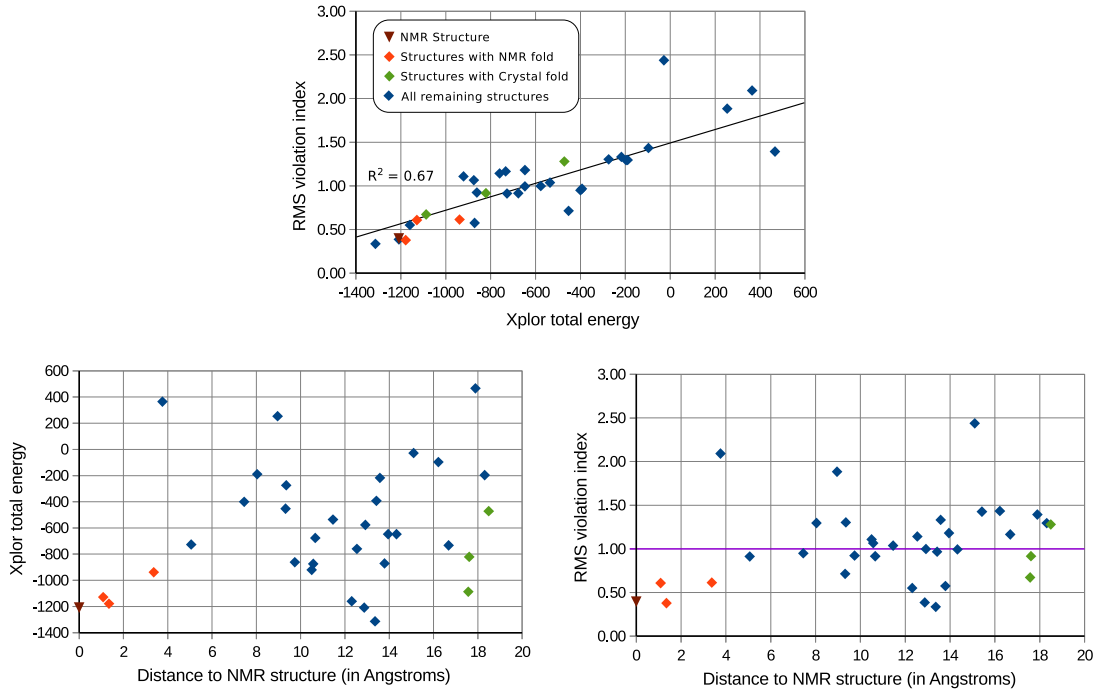


FIGURE 4.6: Top: Structures with low Xplor total energies also have low RMS violation indices. Left: For DAGK, the Xplor total energy function does not have a single low-energy well. Right: The same is true of the RMS violation index, indicating the restraints do not define a unique structure. Structures with a RMS violation index of 1 (purple line) or lower indicate these structures met expectations (on average) for restraint satisfaction. All structural distances are backbone atom (N,C^{α},C′) RMSD values in Å computed for the helical residues 30-48, 51-83, and 90-119 only. Variations in the loop regions were not considered in this score. Even though each structure was refined from a single initial fold, a single fold can describe more than one structure when structures change folds during refinement. For example, two structures changed from their original folds to the NMR fold during refinement, giving the NMR fold three (albeit similar) structures.

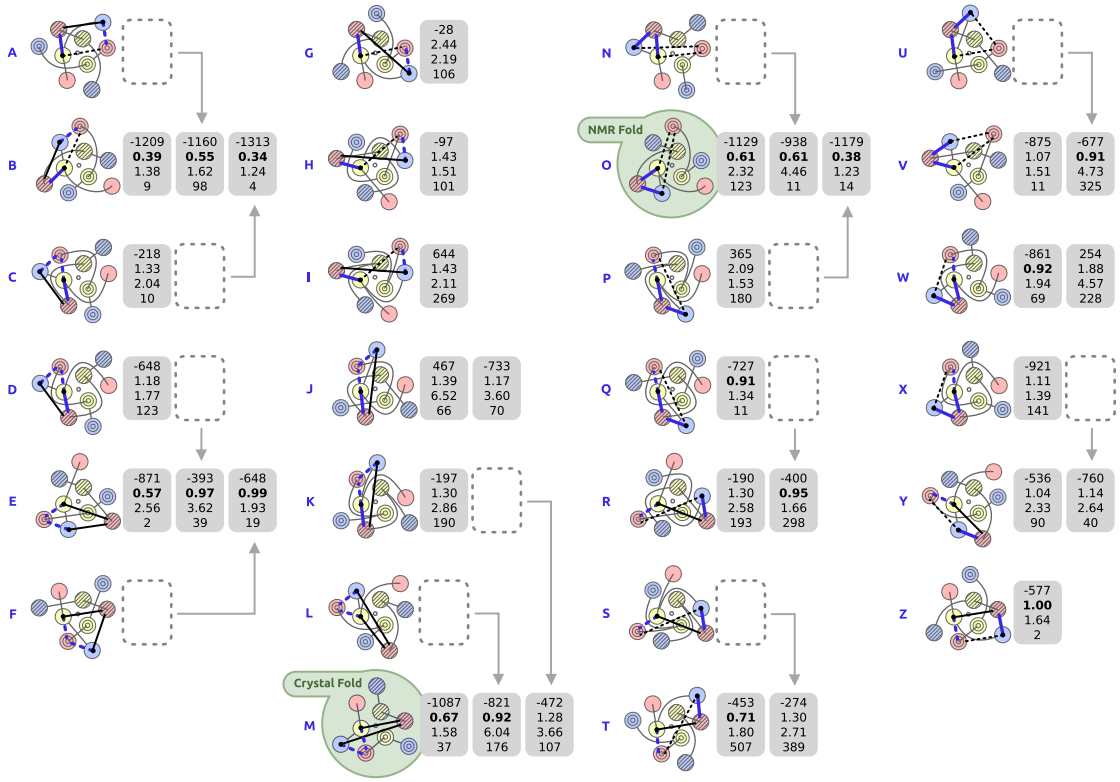


FIGURE 4.7: All 26 DAGK structures grouped by post-refinement fold. Folds are labeled with their names, A-Z, in blue and structures are represented by their statistics in grey boxes. The structure statistics are the same as in Figure 4.5 in the main text. RMS violation indices of 1 or less are highlighted in bold. For structures that switched folds during refinement, an empty box indicates the pre-refinement fold.

5

Bounds for protein backbone dihedral angles from restraints on inter-nuclear vector orientation

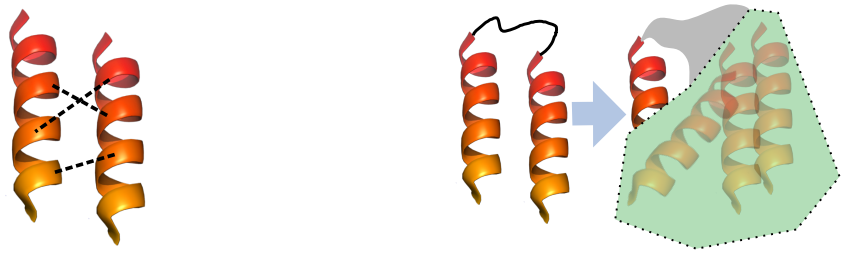
5.1 Introduction

NOE assignment is a major prerequisite to protein structure determination, yet remains a difficult task for spectroscopists since it is plagued by poor experimental sensitivity, chemical shift degeneracy, and also subunit ambiguity. Although DISCO (Chapter 2) can directly address subunit ambiguity and atom ambiguity when analyzing intermolecular NOEs for oligomeric complex structure determination, NOE assignment also poses a significant challenge for monomeric structure determination. Current approaches rely on structural models of homologous proteins (Langmead et al., 2004) or heuristic algorithms that cycle between assignment and structure calculation routines (Herrmann et al., 2002). Structural models are used to assign NOESY spectra, yet the NOE assignments are used to calculate structural models. Convergence to the correct set of assignments, or even convergence at all, is not guaranteed.

The NASCA module from the Donald lab (Zeng et al., 2011) is able to perform

NOE assignment and protein side-chain structure determination simultaneously and with guarantees on solution quality, but requires as input a complete backbone structure of the protein. RDC-ANALYTIC, PACKER, and POOL (Zeng et al., 2009; Tripathy et al., 2011) compute protein backbone structures primarily using restraints from RDCs, but PACKER still requires a few assigned inter-SSE NOEs to pack the SSEs into a core structure. Hence, the algorithms from the Donald lab mitigate the drawbacks of cyclic use of NOEs by merely bootstrapping the rest of the NOE assignments from a few initial assignments (in addition to RDC data), but it would be possible to break the cycle completely if initial NOE assignments were not needed at all. Therefore, we seek a method to compute a complete backbone structure without relying on NOE assignments.

Previous work has computed complete protein backbone structures without any distance restraints from NOEs (Hus et al., 2001; Bryson et al., 2008), but required a large number of restraints from RDC data which may pose a significant challenge to collect for non-model systems. In modern structural studies of challenging protein targets, it is typically possible to find at least one aligning media suitable for the collection of RDC data. Fortunate cases may yield even two or three suitable aligning media. MECANNO (Hus et al., 2001) solved directly for the orientation of each peptide plane in the protein. The peptide plane orientations were then used to assemble the backbone structure of human Ubiquitin, but the calculation required six RDCs per peptide plane in two aligning media. For a model system such as Ubiquitin, an extensive amount of RDC data is available for analysis, including 36 NH RDCs in 18 aligning media (Lange et al., 2008), but there is little hope to collect so much data for a new protein target. RECDRAFT (Bryson et al., 2008) is a more promising candidate for RDC-based backbone structure determination, since it places no strict minimum requirement on RDC data. In practice, however, it is unlikely to efficiently find accurate backbone structures with fewer than two RDC values per peptide plane in



(a) PACKER (b) New approach

FIGURE 5.1: (a): PACKER currently uses NOEs to pack SSEs. Two α-helices are used to illustrate the example, but β-sheets are also applicable. (b): In the new approach, RDCs for the intervening loop are used to pack the SSEs by computing a bound for the position and orientation of the second SSE relative to the first. The bound on SSE placements is then systematically searched to find the optimal packing.

less than two aligning media. In addition, its greedy search algorithm will be unable to guarantee that the computed structures best satisfy the RDC data.

In this chaper, we present work towards extending the Donald lab methodology to remove the requirement of NOE assignments for monomeric protein backbone structure determination, while maintaining guarantees on solution quality. The goal is to provide an alternative to the PACKER module that relies on restraints from RDCs instead of restraints from NOEs (See Figure 5.1). In theory, this algorithm can compute bounds on SSE placement using any number of RDCs in any number of media, but the bounds may only be useful if at least NH and $C^{\alpha}H^{\alpha}$ RDCs in one medium are collected. Once computed, the RDC-based backbone structure can be used by NASCA to simultaneously compute side-chain structures and assign NOESY spectra, and hence solve the complete structure.

The methods described in this chapter do not fully accomplish the goals stated above, but are steps towards a method that one day could. Preliminary results are presented below for an implementation that can compute bounds for extremely idealized information about internuclear vector orientations, but some work remains to improve the implementation to analyze realistic RDC data.

5.2 Applications

Once completed, our method could be used to determine the high-resolution solution NMR structures of two challenging protein targets. The first target is the phenylalanine epimerization domain of the nonribosomal peptide synthetase enzyme gramicidin S synthetase A (Stachelhaus and Marahiel, 1995), or GrsA-PheE, a 57 kDa protein domain consisting of 488 residues. A structural study of GrsA-PheE is currently underway in the Donald lab and the experimental work was conducted by Cheng-Yu Chen. The second target is an N-terminal fragment of Staphylcoccal protein A (SpA-N). The SpA-N construct consists of five nearly identical globular domains connected by flexible linkers. Building on structures of the B domain and related mutants that have been previously solved (Zheng et al., 2004; Gouda et al., 1992; Sato et al., 2004), collaborators Yang Qi and Prof. Terry Oas aim to not only solve the structure of the full construct (where our algorithm could be applicable), but also characterize its dynamics using NMR methodology.

5.3 Results

5.3.1 Bounding protein backbone dihedral angles using RDCs

To compute a bound on the position and orientation of the second SSE relative to the first, we first must compute bounds on the ϕ and ψ torsion angles of the intervening loop backbone. Once computed, the bounds on ϕ and ψ for each loop residue will be analyzed to yield a bound on the position and orientation of the final peptide plane in the loop, which will in turn bound the position and orientation of the second SSE. The method must therefore be tolerant of two types of uncertainty: 1) uncertainty in the experimental RDC measurements and 2) uncertainty in the orientation of a peptide plane. Since the experimentally-measured RDC values will be naturally perturbed by noise, this uncertainty must be tolerated to help ensure

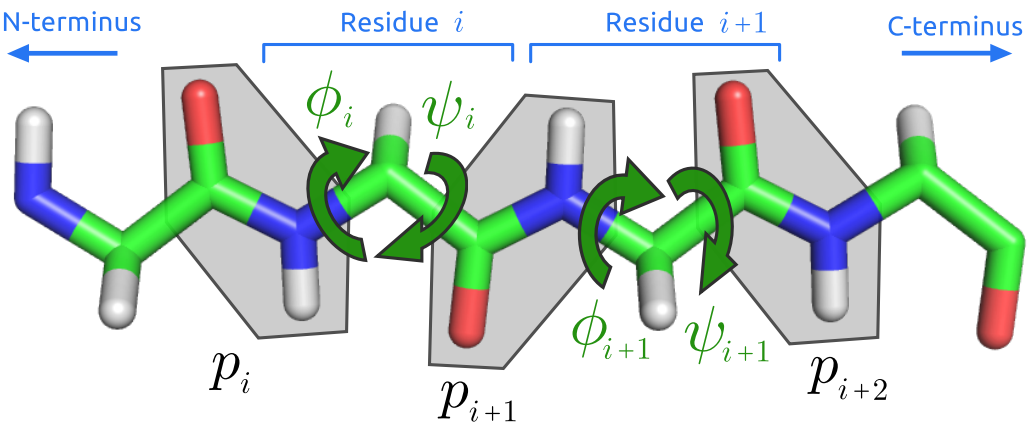


FIGURE 5.2: This fragment of a protein backbone is considered relatively rigid except for dihedral angle degrees of freedom (ϕ, ψ) in each residue. The *peptide planes* are rigid fragments (outlined in grey polygons) that are approximately planar and contain the peptide bonds in the polypeptide. Computing bounds on the dihedral angles begins with an estimate of the orientation of peptide plane p_i. The estimate, along with RDC data are used to compute bounds on ϕ_i and ψ_i. The bounds are then used to compute an estimate of the orientation of p_{i+1}, and so on until the end of the chain is reached.

that all the restraints from RDCs are simultaneously satisfiable. A robust model of uncertainty for the peptide plane orientation is also necessary to enable calculation of the ϕ,ψ bounds inductively along the loop backbone. We assume that each peptide plane along the backbone is completely rigid and its conformation is defined by ideal geometry (Engh and Huber, 1991), hence leaving ϕ and ψ as the only two remaining degrees of freedom for each residue (See Figure 5.2).

The choice of model for the uncertain peptide plane orientations is an important one. Since the source of the orientational restraint originates with RDC data (and hence, sets of internuclear bond orientations), the model must find a way to represent these orientations tightly. For example, such a model might attempt to describe the peptide orientations *explicitly* using subsets of $\mathbb{S}^2 \times \mathbb{S}^1$, a simple proxy for $SO(3)$ (Yershova et al., 2010), where the subset of $\mathbb{S}^2$ is a polar cap bounding the orientation of the NC^α bond vector, and the subset of $\mathbb{S}^1$ is an interval bounding rotations of the peptide plane about the NC^α axis. The drawback to this approach is that the

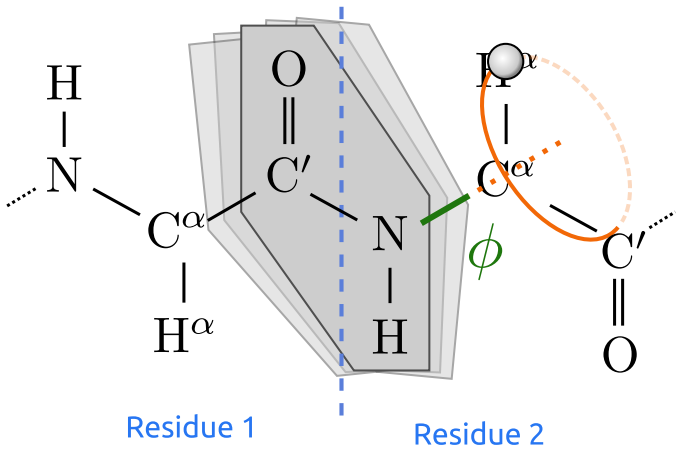


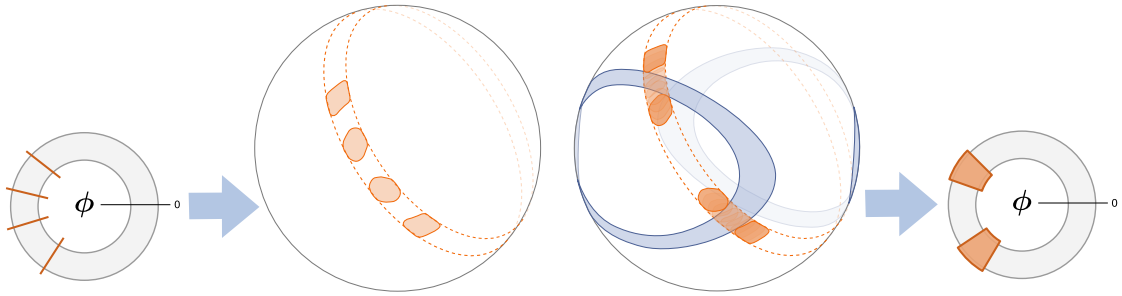
FIGURE 5.3: The ϕ dihedral angle (green) is constrained by measurements of the orientation of the $C^\alpha H^\alpha$ bond vector, which due to forward kinematics, is constrained to lie on a circle (orange). Uncertainty in the orientation of the peptide plane (grey polygons) would "smear" the circle of $C^\alpha H^\alpha$ orientations into a band on the sphere.

subset may describe more rotations than needed (or i.e. be loose) and the size of the ϕ,ψ bounds will grow quickly as the algorithm progresses down the loop backbone. Instead, we have chosen to represent the set of peptide plane orientations *implicitly* using sets of NH and NC^α bond vectors.

The algorithm has three steps, outlined below.

1) Compute bounds for ϕ: Given a set of peptide plane orientations and an RDC value for (e.g.) the $C^\alpha H^\alpha$ bond vector, the goal of step 1 is to compute a bound for the ϕ angle. Assuming the algorithm computes bounds starting with the N-terminal residue of the loop and progresses towards the C-terminal residue, any RDC measurement for a bond vector that is N-wards of the ψ rotatable bond can be used in place of, or in combination with, the $C^\alpha H^\alpha$ RDC. Figures 5.3 and 5.4 shows how the bounds on ϕ are computed from RDC data.

2) Compute bounds for ϕ and ψ: Figure 5.5 shows the method to compute bounds for ϕ and ψ. Like the method used to compute bounds on just ϕ, this method does not

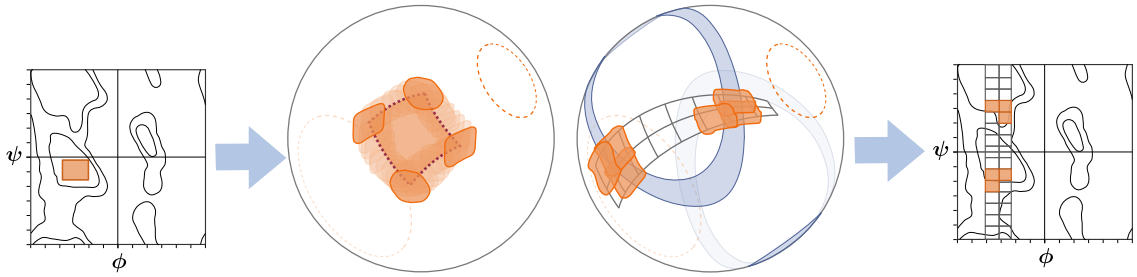


(a) (b)

FIGURE 5.4: (a): Given a set of orientations of the peptide plane N-wards of the current residue, each value for ϕ in $\mathbb{S}^1$ (orange lines, left) describes a corresponding set of $C^\alpha H^\alpha$ bond orientations in $\mathbb{S}^2$ (orange regions, right) that lie in a band on the sphere. (b): Using uncertain RDC data (blue band, left), the goal is to compute all values of ϕ (orange wedges, right) for which the $C^\alpha H^\alpha$ bond orientations (orange regions, left) intersect the RDC band.

require a specific RDC measurement. The examples are illustrated using NH RDCs, but any RDC measurement on the peptide plane C-wards of the current residue can be used.

3) Use the ϕ,ψ bounds to compute the next uncertain peptide plane orientation: After bounds have been computed for ϕ and ψ, the third step is to compute the orientations of the next peptide plane. These orientations are defined by the ϕ,ψ bounds as well as the orientations of the previous peptide plane. The orientations of the NH bond vector are already defined by the RDCs so all that remains to build our implicit model of the peptide plane orientations is to compute the set of orientations for the NC^α bond vector (see Figure 5.6). First, the NH RDCs and rigid peptide geometry define a band B around the sphere which, without any further constraint, describe all possible NC^α bond vectors due to rotation of the peptide plane about the bundle of NH bond vectors. Then, the ϕ,ψ bounds and the set of previous peptide plane orientations define another region R on the sphere of possible NC^α bond vectors similarly to the method described in Figure 5.5a. The final set of NC^α bond vectors



(a) (b)

FIGURE 5.5: (a): Given a set of orientations of the peptide plane N-wards of the current residue, each cell of ϕ,ψ values in $\mathbb{S}^1 \times \mathbb{S}^1$ (orange box, left) describes a set of NH bond orientations in $\mathbb{S}^2$ (orange region, right). (b): Using uncertain RDC data (blue band, left), the goal is to compute all values of ϕ,ψ for which the NH bond orientations intersect the RDC band. As an approximation, we use a hierarchical grid to partition ϕ,ψ space. Each grid cell (clear and orange boxes, right) whose NH bond vector orientations (orange regions, left) intersect the RDC band are labeled satisfying. Satisfying cells are partitioned and their children are recursively analyzed until the desired precision is reached. The final bound on ϕ and ψ is the union of all the satisfying cells.

is the intersection of B and R.

5.3.2 Incremental approach to implementation

Using an implicit model to represent peptide plane orientational uncertainty ensures the orientations described by the RDC data are captured more tightly, but places greater burden on later steps that must interpret these orientations to compute bounds for ϕ and ψ. Due to this additional complexity, we have taken an incremental approach to developing the methodology to implement our ϕ,ψ bounding algorithm (see Figure 5.7).

In the first iteration, for simplicity, we consider a toy version of the ϕ,ψ bounding problem where the peptide plane orientations are represented using a single point for the NH bond vector, and a circular arc for the NC^{α} bond vector. This model allows the peptide plane one degree of rotational freedom about the NH bond vector. Additionally, we assume the constraint imposed by the RDC data on NH and $C^{\alpha}H^{\alpha}$

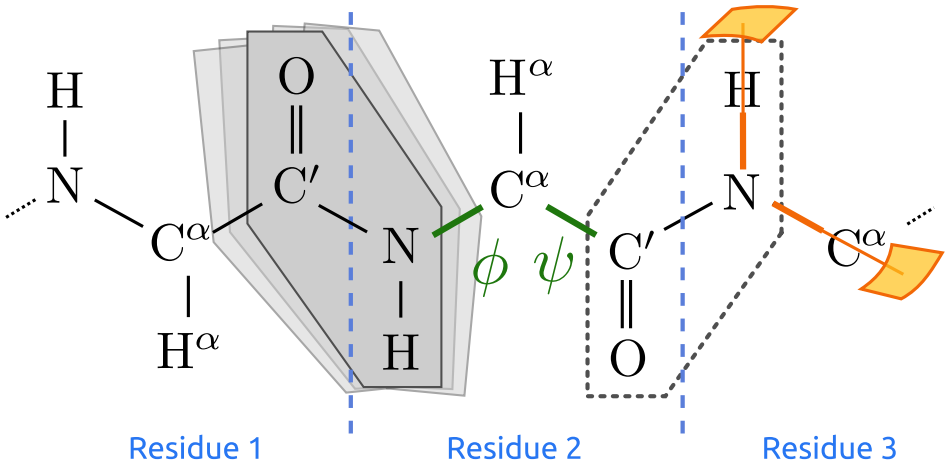


FIGURE 5.6: The set of orientations for the next peptide plane (grey dashed outline) is implicitly defined by the bond vector orientation sets for NH and NC^α (orange regions). Since the NH bond vector orientation is directly defined by RDC data, computing the orientation of the next peptide plane requires using the bounds on ϕ and ψ to compute a set of orientations for the NC^α bond vector.

bond vectors reduces the set of possible orientations for each bond vector to a point on the sphere.

Using this initial model, our preliminary implementation of the ϕ,ψ bounding algorithm was able to quickly reduce orientational uncertainty defined on the initial starting peptide plane to arbitrarily low values for the rest of the polypeptide chain (See Figure 5.8). These results show that, at least under ideal conditions, the ϕ,ψ bounds will not grow uncontrollably as the algorithm progresses down the loop backbone and therefore validates our choice of using the implicit instead of the explicit model for peptide plane rotational uncertainty. The explicit model was tried in early implementations of the algorithm, and the bounds grew uncontrollably for later residues due to the explicit model not being able to tightly capture the orientations described by the bond vector constraints.

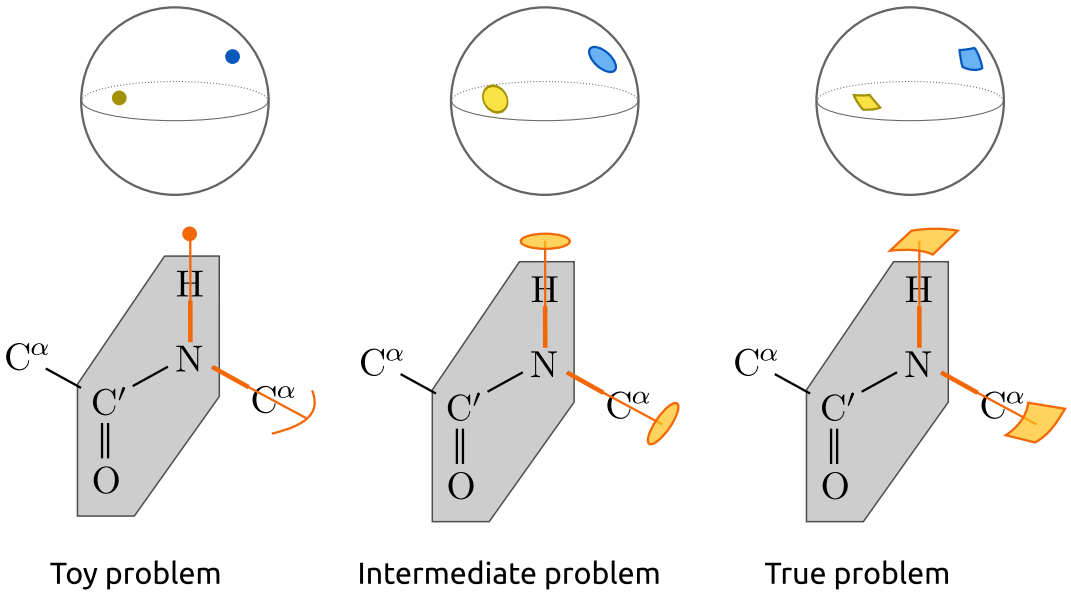


FIGURE 5.7: Models of peptide plane uncertainty and bond vector constraints. Each of the three models uses a set of NH bond vector orientations and a set of NC^{α} bond vector orientations (both orange) to model the orientational uncertainty of the peptide plane (grey polygon). Left: The simplest version models the NH set as a point and the NC^{α} set as a circular arc. This model assumes the RDC data constrain the NH and $C^{\alpha}H^{\alpha}$ bond vector orientations (blue and yellow, respectively) to single points on the sphere. Middle: The second iteration models the NH and NC^{α} sets as regions on the sphere bounded by circular curves (i.e., polar caps). The model assumes RDC data constrains bond vectors to polar caps as well. Right: the final iteration models the NH and NC^{α} sets as regions on the sphere bounded by arcs of RDC curves. RDC curves, also sometimes called sphero-conic curves, are intersections of the unit sphere with a quadric surface. Bond vectors constrained by RDC data are represented by their true intersections of RDC bands, which have the same descriptions as the NH and NC^{α} sets.

5.3.3 A more expressive description for the peptide plane orientational uncertainty

In the second iteration, we used a model for peptide plane orientational uncertainty where the NH and NC^{α} bond vectors were defined by regions on the sphere bounded by circular arcs. The constraint on the NH and NC^{α} bond vectors due to the RDC data is also circular under this model. This model is more practical than the previous iteration since the peptide plane is now allowed some motion in all three orientational degrees of freedom. Also, the RDC constraints represent areas of the sphere instead of points, so this more closely resembles the true version of the problem. When multiple

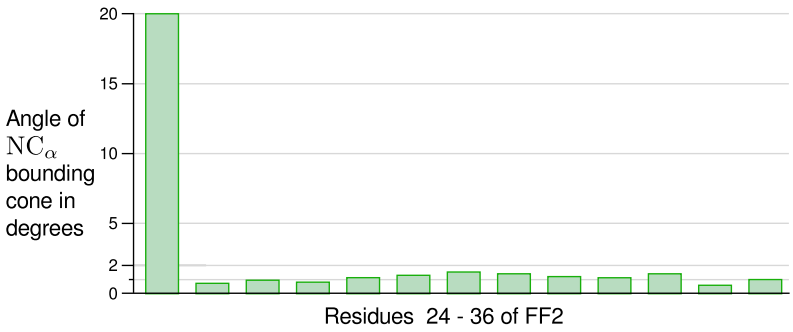


FIGURE 5.8: Size of rotational uncertainty per residue using exact NH and $C^\alpha H^\alpha$ bond orientations from FF Domain 2 of human transcription elongation factor CA150 (Zeng et al., 2009) (FF2). The rotational uncertainty size for each residue was measured by the opening angle of the bounding cone around the C-wards NC^α bond vectors of the peptide planes. The initial uncertainty was set to 20°. The size of the rotational uncertainty of residues 25–36 can be made arbitrarily small by adjusting the precision of approximations in the implementation of the bounding algorithm.

independent RDC datasets constrain bond vectors to small compact regions of the sphere (rather than large bands of constraint defined by just one RDC dataset), then this intermediate model could serve as a simplifying approximation of the true problem (see Figure 5.9). Also, the implementation need only consider the geometry deriving from the analysis of circular arcs and can avoid the difficulty with analyzing the more complicated curves defined by real RDC data (Donald, 2011).

The rest of this chapter will present the geometric details of computing bounds on ϕ and ψ using this more expressive model for peptide plane orientational uncertainty.

5.3.4 An exact bound on the uncertain orientation of the $C^\alpha H^\alpha$ bond vector

First, we describe how to compute a bound on the orientation of the $C^\alpha H^\alpha$ bond given a set of peptide plane orientations described by our intermediate model (see Figure 5.7) and given a single value for the ϕ dihedral angle. Since the orientation of a bond vector is equivalently defined as a point on the unit sphere, we will use these notions interchangeably. Knowing the value of the ϕ dihedral angle means the relative orientation of the $C^\alpha H^\alpha$ bond vector is fixed relative to the peptide plane. Therefore, computing a bound on the $C^\alpha H^\alpha$ orientation involves rigidly propagating

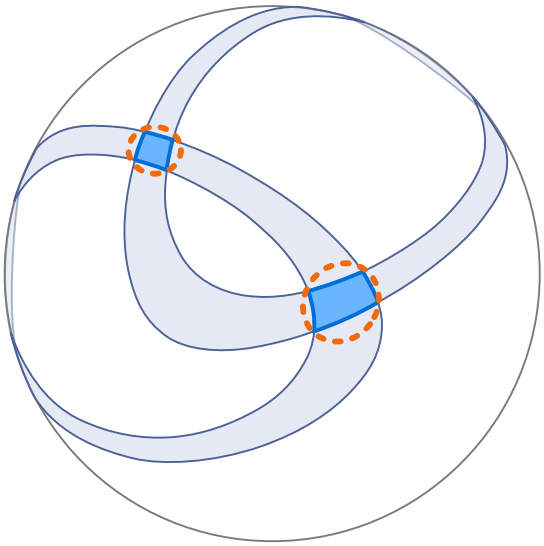


FIGURE 5.9: Bands of constraint defined by multiple RDC datasets (faded blue) sometimes intersect to form smaller regions. These regions (bright blue) are bounded by RDC curves (intersections of quadric surfaces with the sphere) whose geometric complexity complicates analysis, but could be approximated by circular regions (orange).

the orientational uncertainty in the peptide plane to the bond vector orientation. Even though the orientational uncertainty of the peptide plane is described by just circular curves, the bound on the $C^{\alpha}H^{\alpha}$ orientation is remarkably complicated.

One advantage of using circular bounds for the NH and NC^{α} bonds in the peptide plane is the the notion of the "center" orientation is well-defined. Let the center orientations of NH and NC^{α} be $\mathbf{n}$ and $\mathbf{a}$ respectively. We also focus on the case where the size of the two bounds is the same so that both circular regions are the same size. If this were not the case, and one of the regions was defined as smaller than the other, the smaller region would actually invalidate area from the larger region, and consequently the larger region would no longer be circular. Therefore, let the size of these two regions be θ, the opening angle of the circular cones that support the circular curves. Let $C(\mathbf{n}, \theta)$ be a function that returns the region of the unit sphere (and its boundary) enclosed by a circular cone centered at the origin

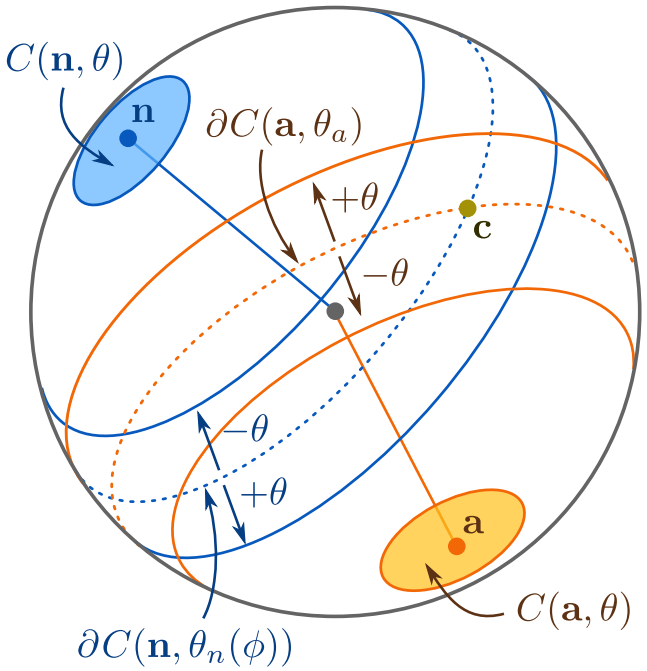


FIGURE 5.10: Part of the bound for the $C^\alpha H^\alpha$ bond vector orientation is due to the circular curves about $\mathbf{n}$ (blue) and $\mathbf{a}$ (orange). The symbols are explained in the text.

with axis $\mathbf{n}$ and opening angle θ. Therefore, the NH and NC^α bounds are denoted $C(\mathbf{n},\theta)$ and $C(\mathbf{a},\theta)$ respectively. The boundary operator ∂ returns just the boundary of these regions, so the boundary of the NH region (i.e., a circular curve) is denoted $\partial C(\mathbf{n},\theta)$. Let $\theta_{n,a}$ be the fixed angle between the NH and NC^α bond vectors in an ideal peptide plane. Due to the rigid conformation of the peptide plane, $\theta_{n,a}$ is always ~120°, so we avoid any degeneracies arising from parallel orientations. Figure 5.10 illustrates this geometry.

We know that if the NH and NC^α bonds in the peptide plane both lie at the centers of their bounding circles, then the $C^\alpha H^\alpha$ bond vector must also lie somewhere inside its bound. Let this orientation of $C^\alpha H^\alpha$ be the "center" of the bound, denoted $\mathbf{c}$. Due to the fixed value of ϕ, we know the angles between the NH, NC^α, and $C^\alpha H^\alpha$ bond vectors are all fixed. This means the $\mathbf{c}$ must lie on the circles $C(\mathbf{n},\theta_n(\phi))$ and $C(\mathbf{a},\theta_a)$, where $\theta_n(\phi)$ is the angle between NH and $C^\alpha H^\alpha$ for a given value of ϕ,

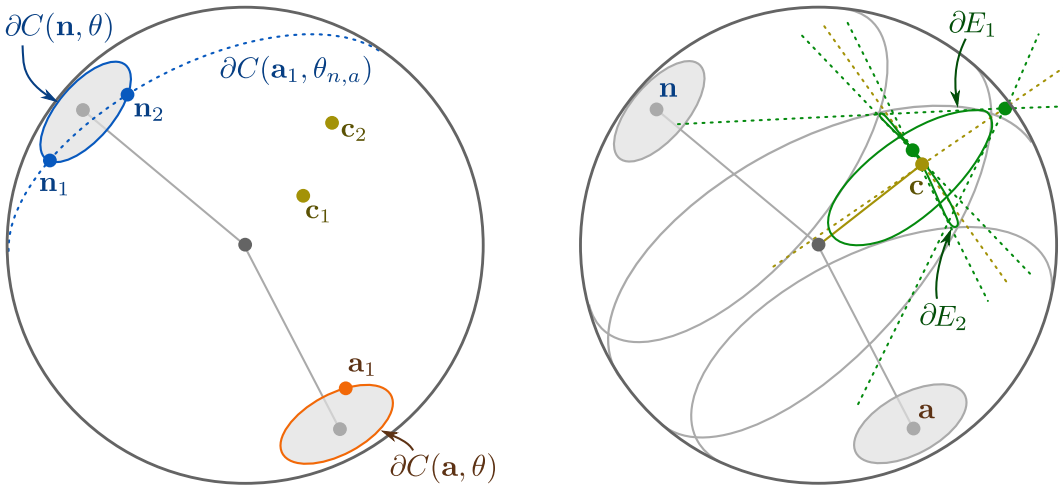


FIGURE 5.11: The second part of the bound for the $C^\alpha H^\alpha$ bond vector orientation. Left: Each sample of $\partial C(\mathbf{a}, \theta)$ gives two samples of $\partial C(\mathbf{n}, \theta)$. Right: The sampled orientations force the NC^α bond to trace two elliptical curves (green). The symbols are explained in the text.

and θ_a is always ~70°. Due to the model of rotational uncertainty for the peptide plane, $C(\mathbf{n}, \theta_n(\phi))$ and $C(\mathbf{a}, \theta_a)$ are "smeared" into circular bands on the sphere (see Figure 5.10). Therefore, the bound for the NC^α bond vector must lie within the intersection of these two circular bands.

We can also learn about the boundary of $C^\alpha H^\alpha$ by analyzing $\partial C(\mathbf{n}, \theta)$ and $\partial C(\mathbf{a}, \theta)$. If we sample a point $\mathbf{a}_1$ from $\partial C(\mathbf{a}, \theta)$, the rigid geometry of the peptide plane requires that orientations of NH lie on $\partial C(\mathbf{a}_1, \theta_{n,a})$. The intersection of $\partial C(\mathbf{a}_1, \theta_{n,a})$ and $\partial C(\mathbf{n}, \theta)$ will always yield one or two points, due to our requirement that the size of the NH and NC^α sets be equal. In the case of two points, let these points be $\mathbf{n}_1$ and $\mathbf{n}_2$. The tuples $(\mathbf{a}_1, \mathbf{n}_1)$ and $(\mathbf{a}_1, \mathbf{n}_2)$ describe two orientations of the peptide plane, and therefore two orientations of the $C^\alpha H^\alpha$ bond, $\mathbf{c}_1$ and $\mathbf{c}_2$ respectively. As the original sample point $\mathbf{a}_1$ moves about $\partial C(\mathbf{a}, \theta)$, $\mathbf{c}_1$ and $\mathbf{c}_2$ trace two different curves on the sphere, ∂E_1 and ∂E_2 respectively. Figure 5.11 illustrates this geometry.

∂E_1 and ∂E_2 each is the curve resulting from the intersection of an elliptical cone

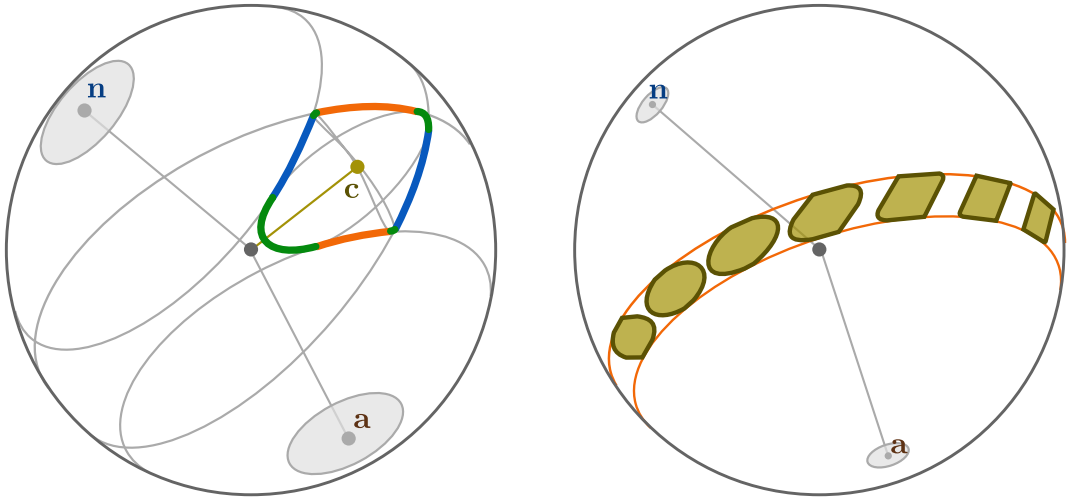


FIGURE 5.12: Left: The boundary of the $C^{\alpha}H^{\alpha}$ bound is constructed using arcs from six different curves. Right: A variety of $C^{\alpha}H^{\alpha}$ bounds from different samples of ϕ. All the bounds lie on the same circular band about $\mathbf{a}$. A smaller value of θ was chosen for this example than on the left.

with the unit sphere, where the apex of each cone is constrained to lie on the sphere, and the axes of both cones pass through $\mathbf{c}$. The apex of each cone results from the projection of $\mathbf{c}$ to another point on the sphere. For one curve, the axis of projection is $\mathbf{n} - \mathbf{a}$. For the other curve, the axis of projection is $\mathbf{n} + \mathbf{a}$. The minor opening angle of the elliptical cones is always $\frac{\theta}{2}$. The relationship of the major opening angles of the cones to the peptide geometry is unknown, but the major opening angle for each elliptical cone can be very precisely fit from samples, which our implementation does in practice.

Once all six curves are computed, $\partial C(\mathbf{n}, \theta_n(\phi) \pm \theta)$, $\partial C(\mathbf{a}, \theta_a \pm \theta)$, ∂E_1, and ∂E_2, the bound of the $C^{\alpha}H^{\alpha}$ bond vector is constructed from the arrangement of these curves (see Figure 5.12). However, since the curves intersect degenerately and are imperfectly defined (the curve parameters themselves are often functions of other complicated geometry and so are represented using floating point numbers for convenience), it is not sufficient to compute the arrangement of the curves to build the bound for the $C^{\alpha}H^{\alpha}$ vector.

5.3.5 Computing degenerate intersection points between imperfectly defined curves

Since these six curves intersect degenerately at a single point per pair of curves, special care is needed to compute the arcs on the boundary of the $C^{\alpha}H^{\alpha}$ bond vector orientation. The vertices on the $C^{\alpha}H^{\alpha}$ boundary are all intersection points between one circular curve and one elliptical curve. Were CGAL able to compute arrangements of polynomial curves on the sphere (currently CGAL only supports plane curves), even that would not be a completely robust approach to computing the degenerate intersection points, since the CGAL is only as precise as its inputs. Minuscule perturbations in the definitions of the six curves could cause an intersection point to be narrowly missed. Instead, we use a specialized method to compute the degenerate intersection point between a circular curve and an elliptical curve.

As mentioned before, the boundary of the intersection of our cone function $C(\mathbf{n}, \theta)$ with the unit sphere produces a circular curve. This same circular curve can also be produced by intersecting the unit sphere with a plane. If we project an elliptical curve onto such a plane, and the circular curve and the elliptical curve intersect degenerately, then the degenerate point of intersection must be a point on the elliptical curve with optimal distance to the center of the circular curve (see Figure 5.13).

To compute the degenerate intersection points between the two curves, we must optimize over the projection of the elliptical curve for distance to the center of the circular curve. Let us translate our projected system so that the center of the circular curve is at the origin of the plane. We therefore want to optimize $|\mathbf{e}|$ where $\mathbf{e}$ is a point on the projected and translated elliptical curve. This optimization reduces to computing the roots of a quartic polynomial, which can be in practice be solved numerically, but our implementation relies on CGAL to compute the roots algebraically. Conversion from algebraic numbers to floating point numbers of course requires iterative numerical methods, but CGAL guarantees extremely high precision

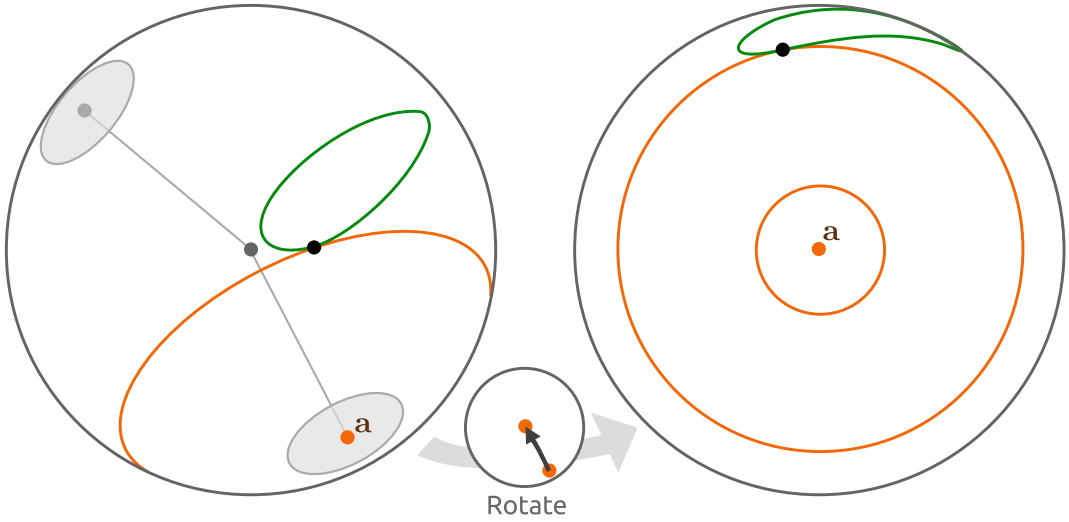


FIGURE 5.13: When viewed in the plane of the circular curve (orange), the elliptical curve (green) intersects at the black point, which has optimal distance to the center of the circular curve.

for this conversion. Once the real roots are computed (up to four), the corresponding points in the plane can be reconstructed and sorted by distance to return the optimal point. Discrimination between the min and max distance points is performed via inclusion predicate with the original circular curve and the points projected back up to the sphere. The inclusion predicate must be tolerant of a small degree of error, since the curves are imperfectly defined.

5.3.6 An exact bound for the certain orientation of the NH bond vector

The conformation of a single reside is specified by a ϕ, ψ pair which resides on the 2-torus, $\mathbb{S}^1 \times \mathbb{S}^1$. To compute a bound on the ψ angle, we must examine a region of ϕ, ψ space and ask whether the NH bond vector from any of these conformations falls within the constrained region (see Figure 5.5). Therefore, our notion of a bound on the NH vector differs slightly from our previous notion of a bound on the $C^{\alpha}H^{\alpha}$ vector. In the $C^{\alpha}H^{\alpha}$ case, we chose a fixed value for the residue conformation (i.e., the ϕ angle) and bounded the $C^{\alpha}H^{\alpha}$ orientation under uncertain orientations of the N-wards peptide plane. In the NH case, there are two scenarios. In the first

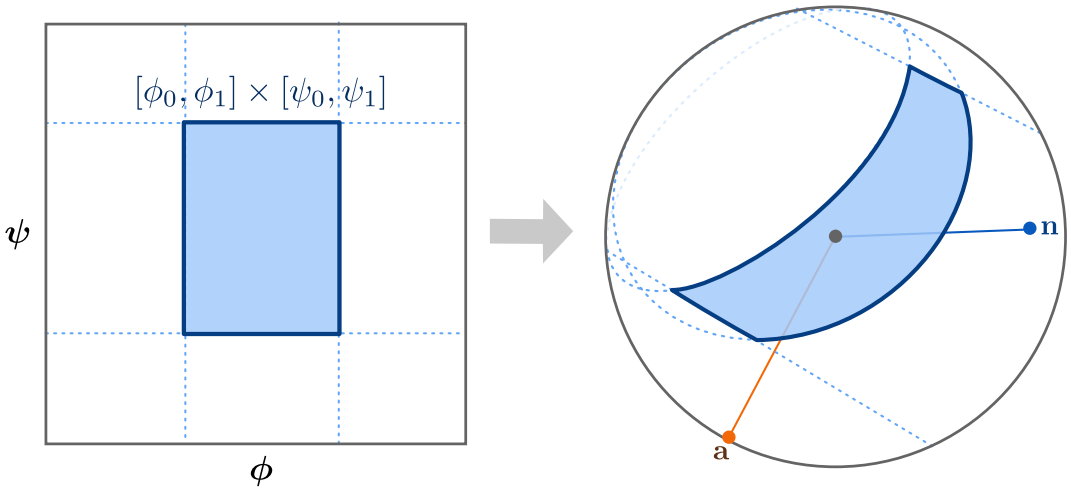


FIGURE 5.14: To describe NH orientations without uncertainty, a ϕ, ψ cell (left) maps to a region on the sphere bounded by circular curves (right).

scenario, we consider the possible orientations of the NH orientation given a set of residue conformations (i.e., ϕ, ψ angles) rather than a single residue conformation, but without any uncertainty in the orientation of the N-wards peptide plane. In the second scenario, we consider the same problem, but with the added peptide plane orientational uncertainty.

For both scenarios, let a ϕ, ψ *cell* be a region of $\mathbb{S}^1 \times \mathbb{S}^1$ defined by $[\phi_0, \phi_1] \times [\psi_0, \psi_1]$. Changes to ϕ and ψ correspond to changes in dihedral angles of the protein backbone, and hence changes to the NH orientation. Without uncertainty, a change to either angle alone causes the NH bond vector to trace a circular curve on the sphere. Therefore, the boundaries of the certain NH orientation induced by a cell are arcs circular curves (see Figure 5.14). Each arc on the boundary of the NH bound is defined by three points sampled from the corresponding edge of the cell.

Remarkably, if all of $\mathbb{S}^1 \times \mathbb{S}^1$ is chosen as the cell, the NC^α orientation of the N-wards peptide plane (and its inverse) always lies outside the bound (see Figure 5.15). This is due to the values of inter-bond angles resulting from the chemistry of organic molecules. The orientation of the NH bond vector avoids the orientations at the

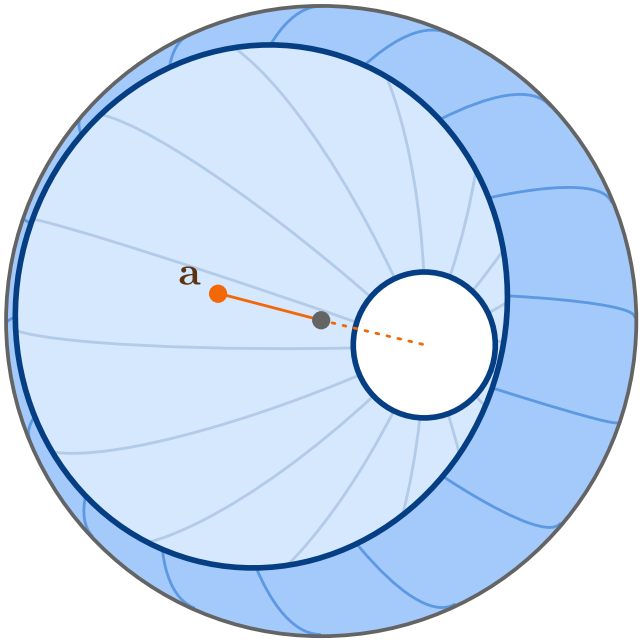


FIGURE 5.15: The range of NH orientations (blue region) accessible by changes in backbone dihedral angles ϕ and ψ remarkably omits the NC^{α} orientation of the N-wards peptide plane (orange line), and its inverse orientation (dashed orange line). Meaning, and NC^{α} orientation can never be parallel to its C-wards NH orientation.

poles of the unit sphere regardless of which values are chosen for ϕ and ψ. Perhaps we are fortunate this is the case, since it means small changes of the NH bond vector orientation (important for its role in stabilizing conformations of proteins via hydrogen bonds) never require large changes in the underlying ϕ, ψ configuration space. Nature has somehow avoided the issues with singularities in this simple map between $\mathbb{S}^1 \times \mathbb{S}^1$ and $\mathbb{S}^2$.

5.3.7 An exact bound for the uncertain orientation of the NH bond vector

Computing a bound on the NH orientations when the orientation of the N-wards peptide plane is uncertain will require tools we developed when computing bounds on the certain NH orientations, when computing bounds on the uncertain $C^{\alpha}H^{\alpha}$ orientations, and also some new tools. The bound on the uncertain NH orientations is similar to the bound on certain NH orientations, but the region for the uncertain

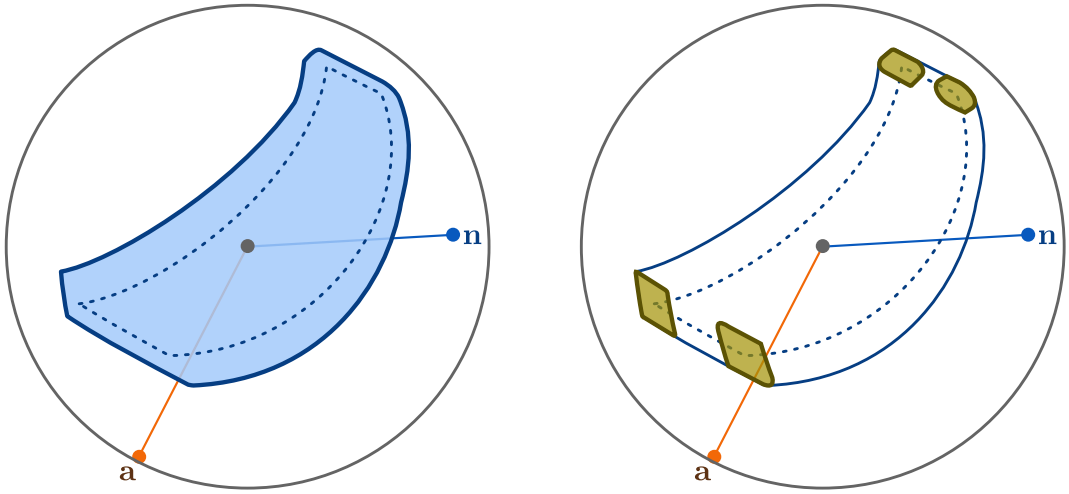


FIGURE 5.16: Left: The certain bound on NH orientations (dashed outline) due to a ϕ, ψ cell is expanded by uncertainty (blue region). Right: Each NH orientation on the boundary of the certain bound (dashed blue outline) was produced by a single residue conformation from the ϕ, ψ cell. Each residue conformation, when coupled with uncertainty, produces an uncertain bound for the NH orientation (yellow). The expansion NH bound due to uncertainty is the envelope of all the single residue conformation uncertain NH orientation bounds (yellow).

orientations is expanded by the uncertainty (see Figure 5.16).

The uncertain bound on NH orientations is bordered three different types of curves (see Figure 5.17): circular curves, elliptical curves, and a third more exotic type of curve. This exotic type of curve can be thought of as an offset curve of a circular curve, but the offset distance is something more complicated than a fixed geodesic distance (see Figure 5.18). The offset is defined by a continuous set of elliptical curves such that the envelope of these elliptical curves is formed from arcs of the exotic offset curve, which we will henceforth refer to as an *elliptical offset curve*. Like many offset curves, this curve also cusps and is not always smooth. We currently do not have an implicit definition of this curve like we do for the circular curves and elliptical curves. Meaning, we know of no supporting surface we can intersect with the sphere to define this curve. At best, we have been able build a parametric representation of this curve which can be used to construct a poly-geodesic approximation with arbitrary precision, and hence simplify intersection

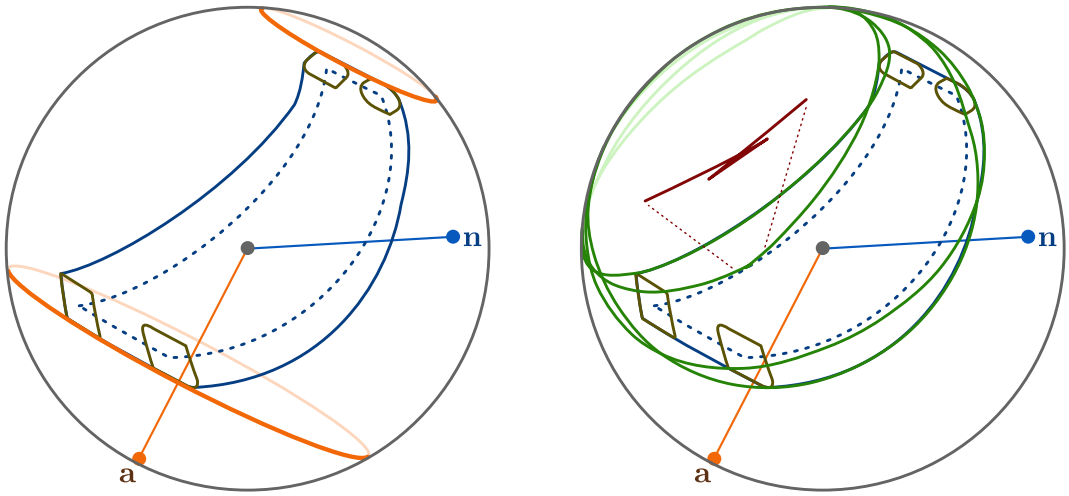


FIGURE 5.17: Types of curves bordering the uncertain NH bound. Left: Circular curves about the N-wards NC^{α} orientation (orange) border part of the NH bound. Curves from the single residue conformation uncertain NH orientation bounds (yellow) support the corners of the bound. Right: More exotic curves border the rest of the NH bound (green). These curves are offset curves and evidence of cusping can be seen even in this very simple example. The red callout magnifies a small section of the green curve to show the cusping in more detail.

calculations with other types of curves.

5.3.8 Parametric description of an elliptical offset curve on the sphere

The definition of the elliptical offset curve derives from geometry of a protein backbone, and our intermediate model of orientational uncertainty for a peptide plane. This curve, being one-dimensional, has one independent parameter $t \in [0, 2\pi)$. Since the elliptical offset curve is defined partially by uncertainty in the peptide plane orientation, each value of t therefore defines an orientation of the peptide plane. This orientation can be described by any two non-parallel vectors in the peptide plane, and we chose the NH and NC^{α} vectors, noted $\mathbf{n}(t)$ and $\mathbf{a}(t)$ respectively. Since our elliptical offset curve is comprised of points from its constituent elliptical curves, the values of $\mathbf{n}(t)$ and $\mathbf{a}(t)$ must lie on the boundaries of their regions, $\partial C(\mathbf{n}, \theta)$ and $\partial C(\mathbf{a}, \theta)$ respectively (see Figure 5.11).

Choosing an orientation of the peptide plane therefore means choosing values for

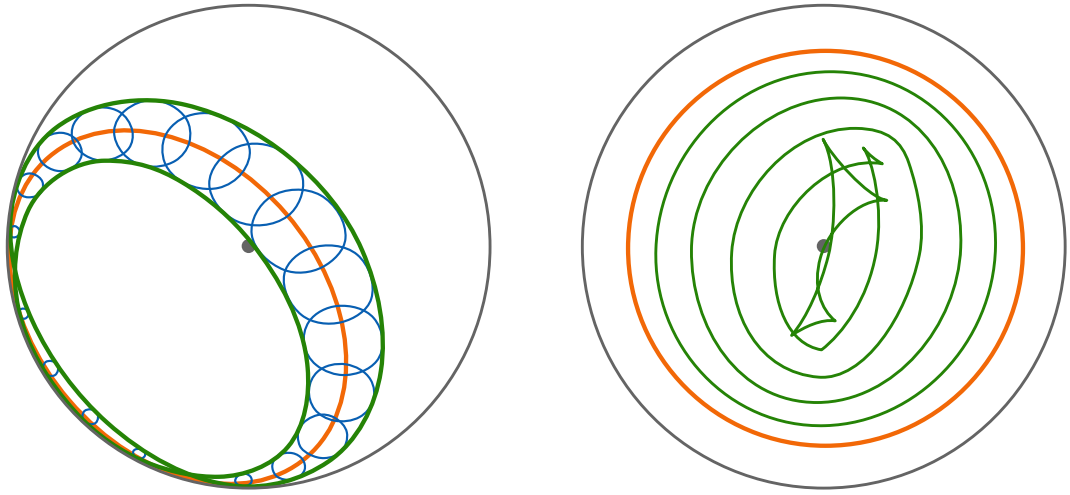


FIGURE 5.18: An elliptical offset curve on the unit sphere. Left: The elliptical offset curves (green) appear to be the envelope of a continuous set of elliptical curves when the elliptical curves are small. The blue curves shown are samples from this continuous set of elliptical curves, which sweep a circular curve (orange) on the sphere. Right: As the sizes of the elliptical cones supporting the elliptical curves grow, the elliptical offset curve (green) also grows and eventually cusps, much like other offset curves. Even though the base curve is circular, the elliptical offset curves grow asymmetrically.

$\mathbf{n}(t)$ and $\mathbf{a}(t)$. If we choose $\mathbf{n}(t)$ arbitrarily from $\partial C(\mathbf{n}, \theta)$, then the value for $\mathbf{a}(t)$ is defined implicitly by the intersection of three surfaces.

$$\mathbf{n}(t) \cdot \mathbf{a}(t) = \cos\theta_{n,a} \tag{5.1}$$

$$\mathbf{a} \cdot \mathbf{a}(t) = \cos\theta \tag{5.2}$$

$$\mathbf{a}(t) \cdot \mathbf{a}(t) = 1 \tag{5.3}$$

This intersection of course yields two points in general. This choice of points defines two related instances of our elliptical offset curve. Choosing one point from this intersection and proceeding with the rest of the derivation will produce one curve. Choosing the other point and proceeding with the derivation will produce a second related curve. The rest of this mathematical description will abstractly describe these intersection points as merely a single point $\mathbf{a}(t)$, but keep in mind that $\mathbf{a}(t)$ is a variable that can be assigned either of the two points from the intersection. For

implementations, some care must be taken to ensure that the same point is chosen for $\mathbf{a}(t)$ when varying the value of t.

Our elliptical offset curve is also defined in part by the ϕ, ψ cell. The two boundary segments of the cell over which ϕ is constant define two values for ϕ. ϕ then becomes a parameter of the elliptical offset curve, but it is not an independent parameter. A value for ϕ defines the orientation of the $C^{\alpha}C$ bond vector relative to its N-wards peptide plane. Therefore, for a given value of ϕ, we can assume the $C^{\alpha}C$ orientation is fixed relative to the peptide plane. Let $\mathbf{v}(t, \phi)$ be the orientation of the $C^{\alpha}C$ bond vector when its N-wards peptide plane is oriented according to $\mathbf{n}(t)$ and $\mathbf{a}(t)$. Again, since the ϕ, ψ cell boundary defines two values of ϕ, these two choices will define two related elliptical offset curves, but the rest of the derivation will treat ϕ as a constant.

Now we have all the tools necessary to define the points on our elliptical offset curve. Let $\mathbf{p}(t)$ be a point on the elliptical offset curve which is implicitly defined by the intersection of three surfaces.

$$\mathbf{v}(t, \phi) \cdot \mathbf{p}(t) = \cos \theta_{p,v} \tag{5.4}$$

$$\left[\frac{\partial}{\partial t}\mathbf{v}(t, \phi)\right] \cdot \mathbf{p}(t) = 0 \tag{5.5}$$

$$\mathbf{p}(t) \cdot \mathbf{p}(t) = 1 \tag{5.6}$$

Here, $\theta_{p,v}$ is the fixed angle between the NH and $C^{\alpha}C$ vectors in an ideal peptide plane, and $\frac{\partial}{\partial t}\mathbf{v}(t, \phi)$ is the vector-valued first derivative of $\mathbf{v}(t, \phi)$ with respect to t which describes the tangent vector at $\mathbf{v}(t, \phi)$.

Intuitively, $\mathbf{p}(t)$ represents an orientation of the NH bond vector in the C-wards peptide plane of the residue. Eq. (5.6) restricts $\mathbf{p}(t)$ to lie on the unit sphere. Eqs. (5.4) and (5.6) define a circular curve on the sphere about $\mathbf{v}(t, \phi)$ which encodes NH orientations allowed by the fixed angle between NH and $C^{\alpha}C$ in a peptide

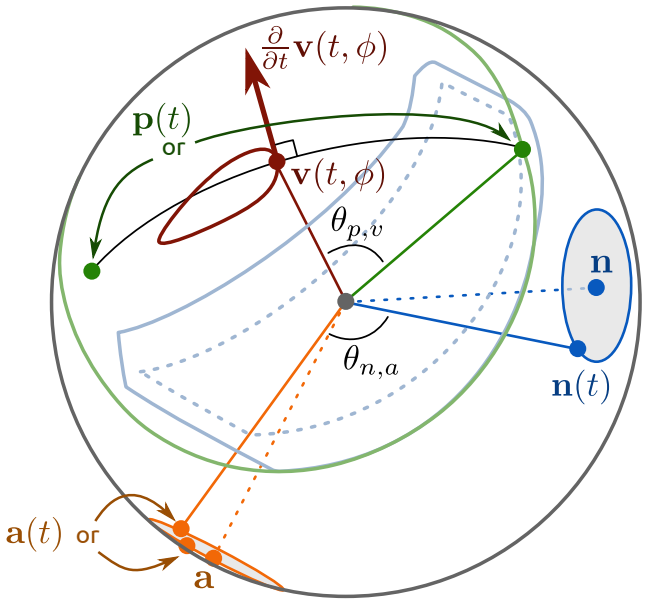


FIGURE 5.19: Supporting geometry of the elliptical offset curve. The symbols are described in the text.

plane. Eqs. (5.5) and (5.6) define a geodesic curve on the sphere that requires NH orientations to be perpendicular to the tangent vector at $\mathbf{v}(t,\phi)$. In effect, these last two constraints enforce that $\mathbf{p}(t)$ has fixed perpendicular geodesic distance from the curve traced by $\mathbf{v}(t,\phi)$. Therefore the elliptical offset curve is like an offset of a circular curve, but that circular curve is also parameterized by t. Figure 5.19 summarizes the geometry used to construct the elliptical offset curves. Another, perhaps simpler, way to think of the elliptical offset curve is a fixed geodesic offset from the elliptical curve traced by the $\mathbf{v}(t,\phi)$ vector. Too bad these things are only obvious only after writing 26 pages of math with pretty pictures...

In general, the intersection described by Eqs. (5.4), (5.5), and (5.6) will again define two points where, again, each choice of point defines a separate curve. Therefore, for a given ϕ,ψ cell, there are eight related elliptical offset curves in total, defined by three separate binary choices.

The base curve of the elliptical offset curve doesn't actually appear explicitly in

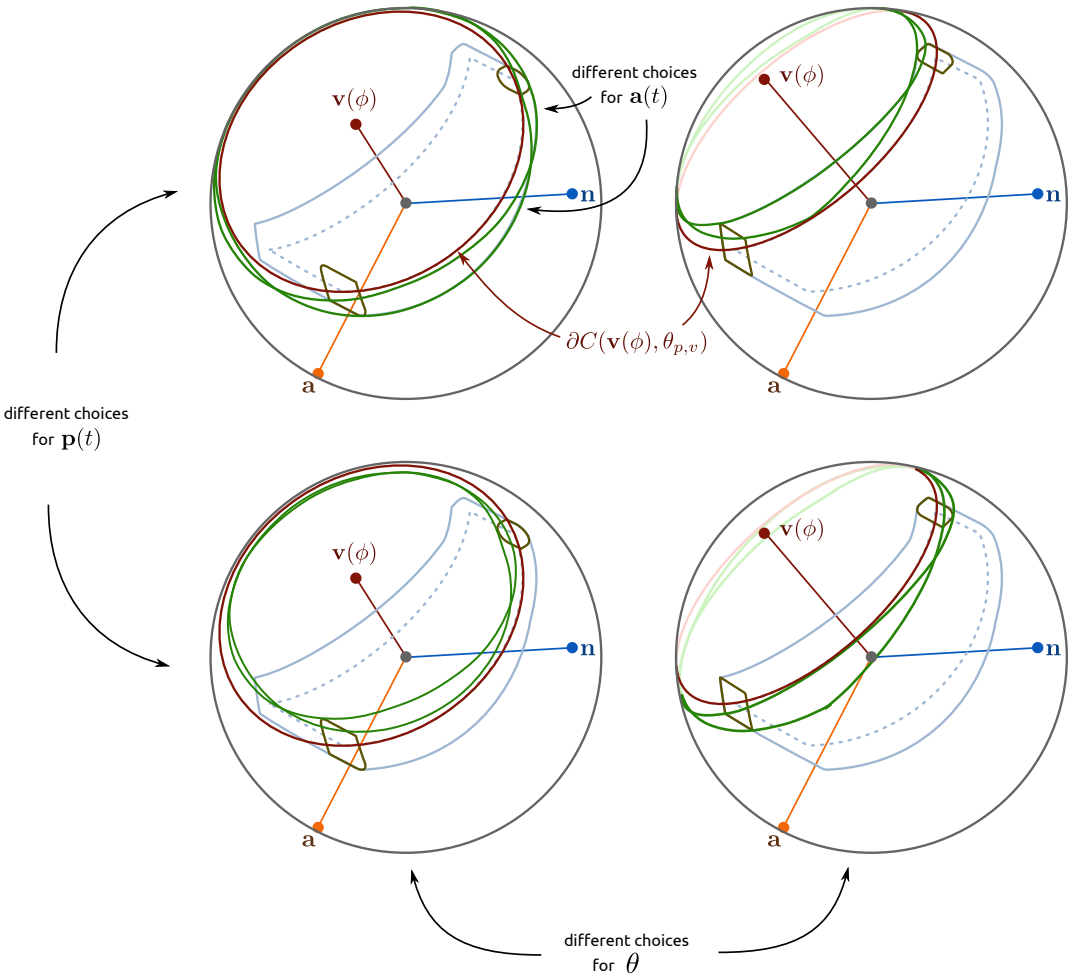


FIGURE 5.20: The family of all eight elliptical offset curves (green) defined by a ϕ, ψ cell. The symbols are described in the text.

its derivation, but we will describe it next. The base curve derives from the $C^{\alpha}C$ vector and is the circular curve $\partial C(\mathbf{v}(\phi), \theta_{p,v})$ where $\mathbf{v}(\phi)$ is the $C^{\alpha}C$ vector when its N-wards peptide plane is oriented according to $\mathbf{n}$ and $\mathbf{a}$. Figure 5.20 shows the family of all eight elliptical offset curves defined by a ϕ, ψ cell.

5.3.9 Future work

The implementation of the algorithm using the intermediate uncertainty model is still unfinished. The implementation can compute bounds on the ϕ angle, but not the ψ angles yet. The exact description of the uncertain bound on NH orientations

is possibly too complex to implement efficiently, and future work should look for conservative approximations. The work done to determine the exact nature of the uncertain NH bound will no doubt be helpful to prove that a later approximate bound is indeed conservative.

The final (and realistic) model for peptide plane orientation uncertainty will directly use the RDC data to define bond vector sets, and it is the most challenging to analyze of the three models. Again, all three degrees of orientational rotation are modeled, but the geometry describing the bond vector sets will be bounded by RDC curves, which are intersections of quadric surfaces with the unit sphere. The additional geometric complexity of the RDC curves must be "pushed through" the algorithm to successfully bound backbone dihedral angles in this case.

A solution to the ϕ,ψ bounding problem with realistic models of peptide plane orientational uncertainty will be able to compute bounds directly from RDC data, and hence will be able to compute bounds on the placements of a second SSE relative to the first SSE using RDCs for the intervening loop backbone. This improvement will allow structure determination methods based on RDC-ANALYTIC and POOL to compute complete backbone structures without NOE assignments. These RDC-defined backbone structures can then be used by NASCA to determine the side-chain structures and also assign NOESY spectra. Being able to compute protein structures with side-chains and assign NOESY spectra without having to rely on homologous protein models or tight assignment/structure calculation loops will further NMR methodology and enable the accurate structure determination of larger and more challenging protein targets.

6

LibProtNMR: A reusable software library for manipulation of protein structures and analysis of NMR data

The Library for Protein NMR (LibProtNmr) is an open-source library of about 35,000 lines of modular and reusable Java code that provides low-level methods useful for implementing algorithms in structural biology. The library is relatively mature, since it has been developed and used over a period of about seven years, and it includes unit tests with wide coverage. Every method in this library was created because it was needed for a research project, so it is likely these methods will be useful to other researchers as well. This chapter outlines the functionality provided by LibProtNmr, describing the main functionality of each module. Where appropriate, the rationale for the design of classes is given, and the ease of invoking these methods is illustrated with short code examples.

LibProtNmr is freely available under the open-source LPGL license. It can be downloaded from the Donald lab website:

`http://www.cs.duke.edu/donaldlab/software.php`

6.1 Protein structure manipulation

Perhaps the first task any computational structural biologist must do is write code to read and write PDB files. For some reason, there were no widely-available libraries in Java to do this when I began my studies, even though the PDB file format has been an enduring standard. Although, at the time of this writing, a recent search has revealed the BioJava library provides the ability to read PDB files, but appears to lack the ability to write PDB files, and therefore may not terribly useful to computational structural biologists or NMR spectroscopists.

LibProtNmr not only reads and writes PDB files (Code sample 1), but it also builds in-memory representations of protein structures that allow for later manipulation and transformation. Protein structures are represented in memory as collections of subunits, residues, and atoms along with indexing structures that allow for fast atom lookups, which are used extensively during NMR data analysis. The library also provides tools to select atom sets from the protein structure, such as residue ranges and atoms along the backbone, tools to apply geometric transformations (such as rotations and translations) to arbitrary sets of atoms, and tools that manage the chemical bond information between atoms, which is not represented in the PDB format. Protein geometry can even be created *de novo* from backbone dihedral angles and idealized peptide planes.

Code sample 1: Read, transform, and write a protein structure.

```
File file = new File( "path/to/structure.pdb" );
Protein structure = new ProteinReader().read( file );
Quaternion q = new Quaternion();
Quaternion.getRotation( q,
        new Vector3( 0, 0, 1 ),
        Math.toRadians( 90 )
);
ProteinGeometry.rotate( structure, q );
new ProteinWriter().write( file, structure );
```

6.2 NMR data processing

LibProtNmr provides methods to read and write a myriad of different restraints from NMR and other experimental methods including NOEs, PREs, hydrogen bonds, disulfide bonds, restraints from TALOS (Cornilescu et al., 1999), scalar couplings, chemical shifts, and orientational restraints from RDCs. NOEs, PREs, and hydrogen bonds are all treated as distance restraints. Restraints from TALOS and scalar couplings are treated as restraints on dihedral angles. Chemical shifts are not necessarily geometric restraints on protein structure, so they are treated in their own category, as are RDC restraints, which are unlike any of the other kinds of restraints.

RDC restraints are so unique, they warrant many special classes and methods for their processing. Since RDC data describes the tumbling of a molecule in solution, it is necessary to characterize this tumbling before the RDC data can be use to analyze protein structure. LibProtNmr provides methods to compute alignment tensors for RDC data that describe this tumbling. Methods are also provided to analyze the alignment tensors and compute magnitude, rhombicity, asymmetry, and the axes and scalings of the principal order frame.

Methods are also provided to compute the agreement of protein structure to all the experimental measurements mentioned above. For distance restraints, these methods compute many metrics including the number of restraint violations, the magnitude of the greatest violation, and RMS deviations from the restraint bounds. The same metrics are computed for dihedral restraints, although the values are reported in angles instead of distances. Again, RDCs are unique. Evaluating protein structure satisfaction of RDC values requires back-computing RDC values from the alignment tensor and and the structure and then comparing them to experimental RDC values. LibProtNmr provides the necessary methods for these comparisons which report RMS deviations of RDC measurements (in Hertz) and also the unitless

Q-factor measure.

6.3 Atom name translation and mapping

PDB files and NMR restraint files describe atoms using names from two different standards. Nomenclature in the PDB changes over time, so software authored in years past may expect atoms described by older standards than the atom names expected by newer software. Since the first step of analyzing NMR data requires mapping restraint definitions to protein structure, it is extremely important to be able to translate between different naming standards. LibProtNmr provides tools to perform these mappings so that the library can interface with other software.

Atom names in PDB files and NMR restraint files are often described using the triplet (subunit name, residue number, atom name) which serves as an address for the atom. While this format allows these files to be easily understood by humans, their use by software requires computationally expensive String comparisons. If only used a few times, these atom lookups do not impose a processing bottleneck, but since NMR data analysis requires so many atom lookups, LibProtNmr translates these atom addresses into a vastly more efficient indexing system. To lookup an atom, the expensive String comparisons and list searches are replaced with constant-time array accesses to speed up computations. Another benefit to address translation is that algorithms using NMR data are freed from performing the error handling associated with mapping restraint definitions to protein structure. This very error-prone step is handled explicitly by dedicated and robust translation methods before invoking data analysis methods.

Finally, NMR restraint files often contain addresses to atoms that do not actually exist in the protein structure. Due to the nature of distance restraints from NOESY data, the unique proton assignment for a restraint is sometimes not known. Instead, the restraints point to imaginary atoms from pseudo-structures (Wüthrich et al.,

1983) as a mechanism to handle the ambiguity of these restraints. LibProtNmr also provides methods to add these pseudo-atoms to protein structures and perform the address lookups so they can be referenced by NMR restraints. Example code performing all of these steps is presented in Code sample 2.

Code sample 2: Read restraints, translate atom names, add pseudo-atoms, and map the restraints to a protein structure

```
// read a protein
Protein protein = new ProteinReader().read( "path/to/protein.pdb" );
NameMapper.ensureProtein( protein, NameScheme.New );

// read some RDCs
List<Rdc<AtomAddressReadable>> rdcsReadable
        = new RdcReader().read( "path/to/RDCs.mr" );
NameMapper.ensureAddresses(
        protein.getSequences(), rdcsReadable, NameScheme.New
);
List<Rdc<AtomAddressInternal>> rdcs
        = RdcMapper.mapReadableToInternal( protein, rdcsReadable );

// read some NOEs with pseudoatoms
List<DistanceRestraint<AtomAddressReadable>> noesReadable
        = new DistanceRestraintReader().read( "path/to/NOEs.mr" );
NameMapper.ensureAddresses(
        protein.getSequences(), noesReadable, NameScheme.New
);
PseudoatomBuilder.getInstance().buildDistanceRestraints(
        protein.getSequences(), noesReadable
);
PseudoatomBuilder.getInstance().build( protein );
List<DistanceRestraint<AtomAddressInternal>> noes
        = DistanceRestraintMapper.mapReadableToInternal(
                noesReadable, protein
        );

// data ready for processing
doAnalysis( protein, rdcs, noes );
```

6.4 Analysis of protein structures and data

LibProtNmr provides a number of methods to evaluate protein structure and NMR data beyond just computing restraint satisfaction. The MolProbity (Chen et al., 2009b) tool probe can be invoked directly from Java code to compute clash scores.

Different conformations of a structure can be optimally aligned and their atoms can be compared using RMSD measures (Code sample 3). LibProtNmr has methods to compute the variance in the position of atoms over a set of structures. There are methods to compute the symmetry axis of a homo-oligomeric protein geometrically from the structure. One can even search over all possible orientations of a protein structure to see if any other orientation satisfies RDCs almost as well as the optimal orientation. Ramachandran statistics of protein structures can be calculated and compared to standard cutoffs. Finally, LibProtNmr has methods to perform clustering of protein structures using atom RMSD as a distance measure.

Code sample 3: Optimal backbone alignment and backbone atom RMSD computation of two conformations of a protein structure.

```
Protein proteinA = getProteinA();
Protein proteinB = getProteinB();
ProteinGeometry.center( proteinA );
StructureAligner.alignOptimallyByAtoms(
        proteinA, proteinB,
        proteinA.backboneAtoms(), proteinB.backboneAtoms()
);
double rmsd = RmsdCalculator.getRmsd(
        proteinA, proteinB,
        proteinA.backboneAtoms(), proteinB.backboneAtoms()
);
```

6.5 Integration with Xplor-NIH

Often, it is desirable to perform energy minimization of a protein structure, to attempt to balance satisfaction of experimental restraints against biophysical knowledge, such as how molecules fill space and interact with their environment. Sometimes, it is only necessary to compute a score of how well a structure fulfills these requirements. LibProtNmr integrates with Xplor-NIH (Schwieters et al., 2006) to perform these tasks. Using process spawning capabilities built into Java, LibProtNmr provides seamless invocations of Xplor-NIH to compute energy functions and even perform energy minimization of structures (Code sample 4). LibProtNmr

sends dynamically generated scripts to the Python interface of Xplor-NIH. Many of the parameters to the Xplor-NIH scripts can be configured via the Java API at run-time, but if more flexibility is needed, the API will accept new templates for the dynamically generated scripts. This integration is particularly useful if structure minimization or energy calculation is an important step within a loop in a structural biological algorithm.

Code sample 4: Integration with Xplor-NIH to perform energy minimization of protein structures.

```
Protein protein = getProtein();
List<DistanceRestraint<AtomAddressReadable>> noes = getNoes();
StructureMinimizer minimizer = new StructureMinimizer();
minimizer.setDistanceRestraints( noes );
minimizer.setNumSteps( 10000 );
Protein minimizedProtein = minimizer.minimize( protein );
```

6.6 Practical geometry and linear algebra

LibProtNmr provides a very simple set of classes to model mathematical and geometrical objects such as vectors, lines, circles, boxes, spheres, annuli, matrices, quaternions, quadratic root solving, etc. The library was designed with memory usage in mind. One of the major bottlenecks to scaling Java programs up to larger input sizes and across multiple threads is managing the memory associated with intermediate calculations. The math libraries in LibProtNmr, as much as possible, rely on the caller to supply allocated memory for computations if such computations cannot be performed using stack memory. Of course, the tradeoff to this approach is that code written to implement math looks less like the original mathematical expression. However, the significant gains in performance, particularly in multi-threaded Java applications where locking global memory allocation structures inside the JVM imposes a significant but unnecessary performance penalty, are a convincing reason to adopt a slightly different way to render math into code. Any readability lost

by using more efficient mathematical libraries can be more than offset by effectively commenting the code. Example code for computing vector expressions is shown in Code sample 5.

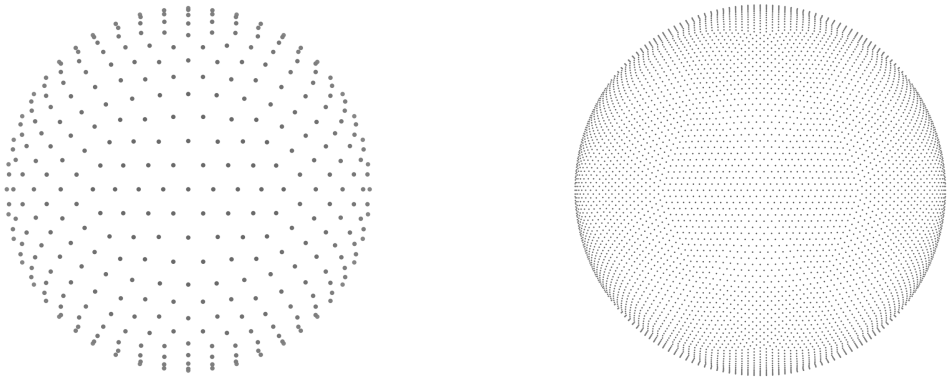
Code sample 5: Vector manipulation using LibProtNmr

```
Vector3 v = new Vector3( 1, 2, 3 ); // two memory
Vector3 w = new Vector3( 4, 5, 6 ); // allocations

// compute |-3( v/|v| + w )|
// no additional object memory needed
v.normalize();
v.add( w );
v.negate();
v.scale( 3 );
double result = v.getLength();
```

Sampling the sphere is a common task that occurs in processing of geometric data. While completely uniform sampling of the sphere is possible, often a quick and dirty approximation to uniform spherical sampling is all that is really needed for a given application. LibProtNmr implements a near-uniform sampling of the sphere using a multi-resolution hierarchical grid of geodesic arcs created from subdividing the faces of a regular icosahedron. This approach allows the caller to select the desired sampling resolution, the computation is quick, and the results are good enough when only approximations are needed (see Figure 6.1). Computing the min bounding sphere from a set of points is also supported via an implementation of Welzl's algorithm (Welzl, 1991).

For more advanced linear algebra (often used in the analysis of RDC data), LibProtNmr integrates with the Jama library to perform these computations, including principal component analysis, eigenvalue/spectral decomposition, singular value decomposition, and QR factorization. Numerical error resulting from the usage of LibProtNmr's geometry and mathematical libraries is handled in a rudimentary way by using epsilon-based comparison of floating point numbers. However, when exact precision is actually needed, critical methods can be implemented using multi-



(a) 3 subdivisions (b) 5 subdivisions

FIGURE 6.1: Fast near-uniform sampling of the sphere using icosahedral approximations.

precision floating point numbers, or exact number types using the computational geometry algorithms library (CGAL).

6.7 Visualization using KiNG

When working with geometric data, visualizing the results of computations in three-dimensions (even intermediate computations) can often provide much more insight than tables, plots, and debuggers. For this reason, many of the geometrical objects provided by LIBPROTNMR can be rendered into a Kinemage for use with the 3D display software KiNG (Chen et al., 2009a). Figures 6.2 and 6.3 show examples of the kinds of geometry that can be rendered with LIBPROTNMR and KiNG. Example code for creating Kinemages is presented in Code sample 6.

6.8 Plotting

While KiNG is great for interacting with 3D graphics, it is not terribly great for rendering publication-quality figures. For this, LIBPROTNMR provides a plotting module that builds on the popular open source JFREECHART library. Plotting is limited to 2D displays of information, so it cannot render protein structures, but PY-

Code sample 6: Display 3D geometry using KiNG.

```
// render the sphere samples from Figure 6.1
GeodesicGrid grid = new GeodesicGrid( 5 );
Kinemage kin = new Kinemage();
KinemageBuilder.appendPoints( kin, grid.vertices() );
new KinemageWriter().show( kin );

// show a protein structure, and wait for the user
Protein protein = getProteinFromSomewhere();
kin = new Kinemage();
KinemageBuilder.appendAxes( kin );
KinemageBuilder.appendProtein( kin, protein );
new KinemageWriter().showAndWait( kin );
```

Mol (Schrödinger, 2012) is the typical choice for that task anyway. LibProtNmr's plotting module can plot various specialized graphs though, including comparisons of experimental vs back-computed RDC values (Code sample 7), RDC histograms, representations of functions over the sphere (Figure 6.4), geometry in the plane (Figure 2.5), even geometry in phi,psi space like Ramachandran statistics.

6.9 Utilities

LibProtNmr has a number of utility classes that perform tasks not necessarily related to structural biology, but nonetheless are useful tools for building software that performs structural biological computations. There are many tools that fall into this category, but three of them are widely used and applicable to many situations: the timer, the profiler, and the progress bar. Since structural biological algorithms often perform very sophisticated computations, these computations can take a lot of wall clock time. These three tools are designed to help manage software performance and also the time of the computational scientist. The timer does just what you think it should. It measures wall clock time between defined start and end lines of code (See Code sample 8). The profiler is also very straightforward. Of course, many full-featured tools for profiling software exist, but sometimes the simple tools are the most useful. The profiler in LibProtNmr works in much the same way the

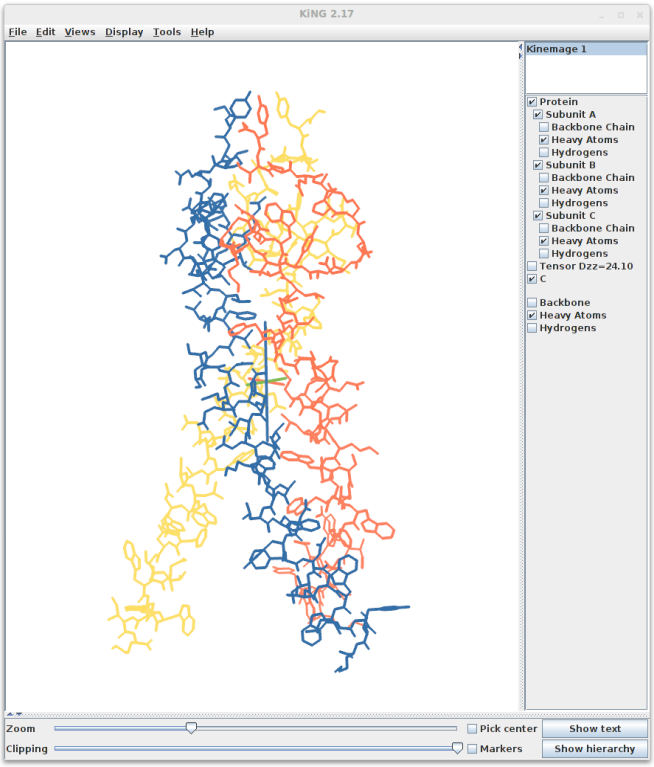


FIGURE 6.2: a protein structure rendered by LIBPROTNMR in KiNG along with the principal axes of the alignment.

timer does. Start and end lines of code are defined, and the wall clock time between them is measured. Where it differs from the timer though, is that each period of time is collected into user-defined bins and the final aggregate running times can be reported (See Code sample 9). This allows the computational scientist to quickly narrow down on performance bottlenecks in the software just by writing a few lines of code without having to rely on complicated or cumbersome profiling frameworks.

Perhaps the most useful of the three tools is the progress bar (Code sample 10). The progress bar is initialized with a number of units of work to be performed. Then during execution of the software, the progress bar is updated with the number of units of work completed. Not only can the progress bar then periodically report what percentage of the work has been completed, but it also attempts to estimate the amount of wall clock time remaining until the end of the task. Since both linear and

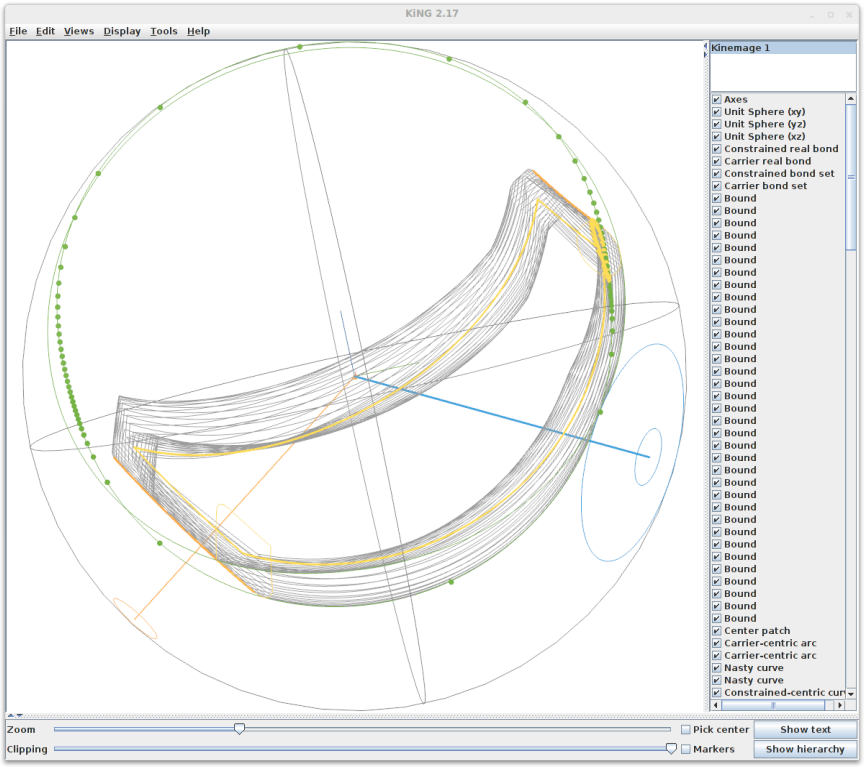


FIGURE 6.3: Viewing complicated abstract geometry is also possible using KiNG.

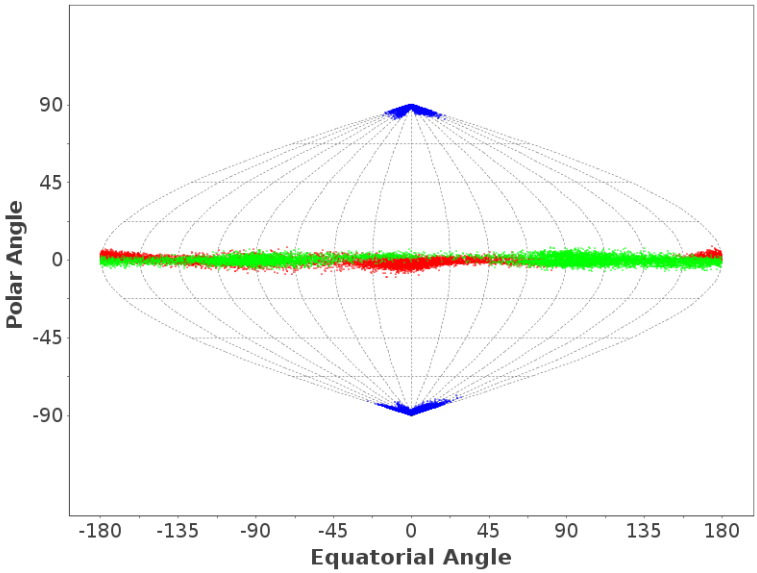
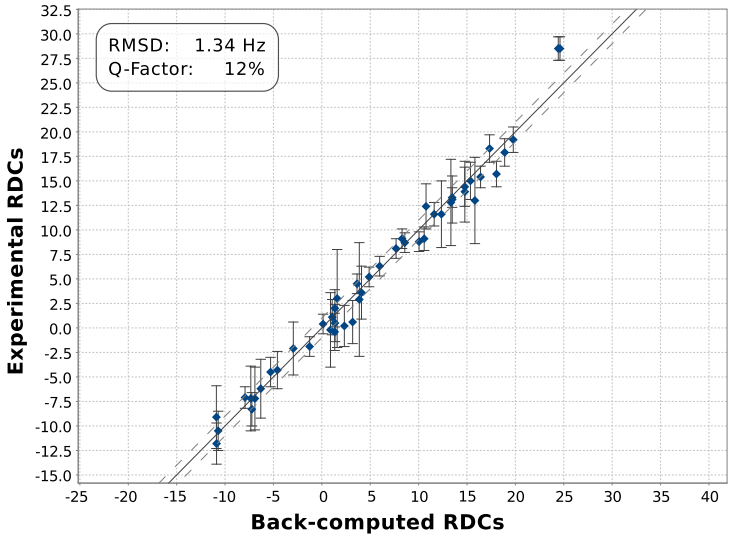


FIGURE 6.4: LIBPROTNMR can plot functions over the sphere using a Sanson-Flamsteed projection (Bugayevskiy and Snyder, 1995).

Code sample 7: Rendering specialized plots takes only a few lines of code.

```
Protein protein = getProtein();
List<Rdc<AtomAddressInternal>> rdcs = getRdcs();
AlignmentTensor tensor = AlignmentTensor.compute(
        protein, rdcs
);
ChartWriter.showAndWait(
        Plotter.plotRdcFit( rdcs, tensor, protein )
);
```



Code sample 8: Timers provide a simple way to report the running time of algorithms.

```
Timer timer = new Timer();
timer.start();
doSomethingComplicated();
System.out.println( timer.getElapsedTime() );
```

quadratic prediction models are available to the progress bar, it can fairly accurately predict the time remaining for a wide range of tasks. This tool lets the computational scientist decide shortly after the start of a long computation whether or not it would be worth the time to wait for the results.

6.10 Python bindings

All of the functionality in LibProtNmr is accessible to Python scripts as well as Java applications thanks to the JPype compatibility layer. The flexibility of a

Code sample 9: Profilers are a useful and simple tool to find performance bottlenecks.

```
Profiler.start( "total" );
while( condition )
{
        doSomethingUnimportant();
        Profiler.start( "important" );
        doSomethingImportant();
        Profiler.stop( "important" );
        doSomethingUnimportant();
}
Profiler.stop( "total" );
System.out.println( Profiler.getReport() );
System.out.println( Profiler.getMemoryUsed() );
```

Code sample 10: Progress bars show long one might have to wait for a result.

```
Progress progress = new Progress( 1000, 5000 );
for( int i=0; i<1000; i++ )
{
        doWork();
        progress.incrementProgress();
        // running time estimates are automatically
        // written to stdout every 5000 ms
        // a final report is written at 100%
}

// if an algorithm has quadratic running time,
// a different prediction model can be used
progress = new Progress( 1000, 5000, Model.Quadratic );
for( int i=0; i<1000; i++ )
{
        doQuadraticWork( i );
        progress.incrementProgress();
}
```

scripting environment allows many separate or short tasks to be completed using LibProtNmr when writing a full-blown Java application would be overkill (Code sample 11).

Code sample 11: Python bindings make LibProtNmr's catalog of modules available in a scripting environment.

```
import jvm, libprotnmr

# start the jvm
jvmInstance = jvm.Jvm()
jvmInstance.addPath( "libprotnmr.jar" )
libprotnmr.initJvm( jvmInstance )
jvmInstance.start( "-Xmx512m" )

# call libprotnmr functions for common tasks
protein = libprotnmr.loadProtein( "path/to/protein.pdb" )
rdcs = libprotnmr.loadRdcs( "path/to/rdcs.mr" )
rdcs = libprotnmr.mapRdcsToProtein( rdcs, protein )

# import Java class names to use other modules directly
AlignmentTensor = libprotnmr.f.nmr.AlignmentTensor

# compute an RDC fit
tensor = AlignmentTensor.compute( protein, rdcs )
print "RDC Q-factor: %.1f"
        % tensor.getQFactor( protein, rdcs )*100
```

Bibliography

Al-Hashimi, H. M., Bolon, P. J., and Prestegard, J. H. (2000), "Molecular Symmetry as an Aid to Geometry Determination in Ligand Protein Complexes," *Journal of Magnetic Resonance*, 142, 153–158.

Alam, S. M., Scearce, R. M., Parks, R. J., Plonk, K., Plonk, S. G., Sutherland, L. L., Gorny, M. K., Zolla-Pazner, S., VanLeeuwen, S., Moody, M. A., et al. (2008), "Human immunodeficiency virus type 1 gp41 antibodies that mask membrane proximal region epitopes: antibody binding kinetics, induction, and potential for regulation in acute infection," *Journal of virology*, 82, 115–125.

Alam, S. M., Morelli, M., Dennison, S. M., Liao, H.-X., Zhang, R., Xia, S.-M., Rits-Volloch, S., Sun, L., Harrison, S. C., Haynes, B. F., et al. (2009), "Role of HIV membrane in neutralization by two broadly neutralizing antibodies," *Proceedings of the National Academy of Sciences*, 106, 20234–20239.

Arora, A. (2013), "Solution NMR spectroscopy for the determination of structures of membrane proteins in a lipid environment." *Methods Mol. Biol.*, 974, 389–413.

Ashkenazi, A. and Shai, Y. (2011), "Insights into the mechanism of HIV-1 envelope induced membrane fusion as revealed by its inhibitory peptides," *European Biophysics Journal*, 40, 349–357.

Bardiaux, B., Bernard, A., Rieping, W., Habeck, M., Malliavin, T. E., and Nilges, M. (2009), "Influence of different assignment conditions on the determination of symmetric homodimeric structures with ARIA," *Proteins: Structure, Function, and Bioinformatics*, 75, 569–585.

Bartesaghi, A., Merk, A., Borgnia, M. J., Milne, J. L., and Subramaniam, S. (2013), "Prefusion structure of trimeric HIV-1 envelope glycoprotein determined by cryo-electron microscopy," *Nature structural & molecular biology*, 20, 1352–1357.

Bellot, G., McClintock, M. A., Chou, J. J., and Shih, W. M. (2013), "DNA nanotubes for NMR structure determination of membrane proteins." *Nat Protoc*, 8, 755–70.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000), "The Protein Data Bank," *Nucl. Acids Res.*, 28, 235–242.

Bryson, M., Tian, F., Prestegard, J. H., and Valafar, H. (2008), "REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data." *J. Magn. Reson.*, 191, 322–34.

Bugayevskiy, L. M. and Snyder, J. P. (1995), *Map Projections: A Reference Manual*, Taylor & Francis, London/Bristol.

Burton, D. R. (2010), "Scaffolding to build a rational vaccine design strategy," *Proceedings of the National Academy of Sciences*, 107, 17859–17860.

Burton, D. R., Desrosiers, R. C., Doms, R. W., Koff, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. (2004), "HIV vaccine design and the neutralizing antibody problem," *Nature immunology*, 5, 233–236.

Bushman, F. D., Engelman, A., Palmer, I., Wingfield, P., and Craigie, R. (1993), "Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding." *Proc. Natl. Acad. Sci. U.S.A.*, 90, 3428–32.

Buzon, V., Natrajan, G., Schibli, D., Campelo, F., Kozlov, M. M., and Weissenhorn, W. (2010), "Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions." *PLoS Pathog.*, 6, e1000880.

Byeon, I.-J. L., Louis, J. M., and Gronenborn, A. M. (2003), "A Protein Contortionist: Core Mutations of GB1 that Induce Dimerization and Domain Swapping," *Journal of Molecular Biology*, 333, 141–152.

Cardoso, R. M., Zwick, M. B., Stanfield, R. L., Kunert, R., Binley, J. M., Katinger, H., Burton, D. R., and Wilson, I. A. (2005), "Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41," *Immunity*, 22, 163–173.

Chan, D. C., Fass, D., Berger, J. M., and Kim, P. S. (1997), "Core structure of gp41 from the HIV envelope glycoprotein." *Cell*, 89, 263–73.

Chen, V. B., Davis, I. W., and Richardson, D. C. (2009a), "KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program," *Protein Science*, 18, 2403–2409.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2009b), "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, 66, 12–21.

Chen, Y., Zhang, J., Hwang, K.-K., Bouton-Verville, H., Xia, S.-M., Newman, A., Ouyang, Y.-B., Haynes, B. F., and Verkoczy, L. (2013), "Common tolerance mechanisms, but distinct cross-reactivities associated with gp41 and lipids, limit production of HIV-1 broad neutralizing antibodies 2F5 and 4E10," *The Journal of Immunology*, 191, 1260–1275.

Clore, G. M., Gronenborn, A. M., and Tjandra, N. (1998), "Direct Structure Refinement against Residual Dipolar Couplings in the Presence of Rhombicity of Unknown Magnitude," *Journal of Magnetic Resonance*, 131, 159–162.

Coggins, B. E. and Zhou, P. (2003), "PACES: Protein sequential assignment by computer-assisted exhaustive search." *J. Biomol. NMR*, 26, 93–111.

Coggins, B. E., Venters, R. A., and Zhou, P. (2010), "Radial sampling for fast NMR: Concepts and practices over three decades." *Prog Nucl Magn Reson Spectrosc*, 57, 381–419.

Cohen, F. E., Sternberg, M. J., and Taylor, W. R. (1980), "Analysis and prediction of protein beta-sheet structures by a combinatorial approach." *Nature*, 285, 378–82.

Cornilescu, G., Delaglio, F., and Bax, A. (1999), "Protein backbone angle restraints from searching a database for chemical shift and sequence homology," *Journal of biomolecular NMR*, 13, 289–302.

Coutant, J., Yu, H., Clément, M.-J., Alfsen, A., Toma, F., Curmi, P. A., and Bomsel, M. (2008), "Both lipid environment and pH are critical for determining physiological solution structure of 3-D-conserved epitopes of the HIV-1 gp41-MPER peptide P1," *The FASEB Journal*, 22, 4338–4351.

DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., and Baker, D. (2009), "Refinement of protein structures into low-resolution density maps using rosetta." *J. Mol. Biol.*, 392, 181–90.

Donald, B. R. (2011), *Algorithms in Structural Molecular Biology*, MIT Press, Cambridge, MA.

Donald, B. R. and Martin, J. (2009), "Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints," *Progress in Nuclear Magnetic Resonance Spectroscopy*, 55, 101–127.

Doyle-Cooper, C., Hudson, K. E., Cooper, A. B., Ota, T., Skog, P., Dawson, P. E., Zwick, M. B., Schief, W. R., Burton, D. R., and Nemazee, D. (2013), "Immune tolerance negatively regulates B cells in knock-in mice expressing broadly neutralizing HIV antibody 4E10," *The Journal of Immunology*, 191, 3186–3191.

Engh, R. A. and Huber, R. (1991), "Accurate bond and angle parameters for X-ray protein structure refinement," *Acta Crystallographica Section A: Foundations of Crystallography*, 47, 392–400.

Fischer, M. W. F., Losonczi, J. A., Weaver, J. L., and Prestegard, J. H. (1999), "Domain Orientation and Dynamics in Multidomain Proteins from Residual Dipolar Couplings†," *Biochemistry*, 38, 9013–9022.

Freed, E. O. (2001), "HIV-1 replication," *Somatic cell and molecular genetics*, 26, 13–33.

Frey, G., Peng, H., Rits-Volloch, S., Morelli, M., Cheng, Y., and Chen, B. (2008), "A fusion-intermediate state of HIV-1 gp41 targeted by broadly neutralizing antibodies," *Proceedings of the National Academy of Sciences*, 105, 3739–3744.

Gallo, S. A., Finnegan, C. M., Viard, M., Raviv, Y., Dimitrov, A., Rawat, S. S., Puri, A., Durell, S., and Blumenthal, R. (2003), "The HIV Env-mediated fusion reaction," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1614, 36–50.

Gautier, A. (2013), "Structure determination of α-helical membrane proteins by solution-state NMR: Emphasis on retinal proteins." *Biochim. Biophys. Acta*.

Gilbert, P. B., McKeague, I. W., Eisen, G., Mullins, C., Gueye-NDiaye, A., Mboup, S., and Kanki, P. J. (2003), "Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal," *Stat Med*, 22, 573–593.

Goodsell, D. S. and Olson, A. J. (2000), "Structural Symmetry and Protein Function," *Annual Review of Biophysics and Biomolecular Structure*, 29, 105–105, doi:10.1146/annurev.biophys.29.1.105.

Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y., and Shimada, I. (1992), "Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures," *Biochemistry*, 31, 9665–9672.

Gobl, C., Dulle, M., Hohlweg, W., Grossauer, J., Falsone, S. F., Glatter, O., and Zangger, K. (2010), "Influence of phosphocholine alkyl chain length on peptide-micelle interactions and micellar size and shape," *The Journal of Physical Chemistry B*, 114, 4717–4724.

Guenaga, J., Dosenovic, P., Ofek, G., Baker, D., Schief, W. R., Kwong, P. D., Hedestam, G. B. K., and Wyatt, R. T. (2011), "Heterologous Epitope-Scaffold Prime Boosting Immuno-Focuses B Cell Responses to the HIV-1 gp41 2F5 Neutralization Determinant," *PLoS One*, 6, e16074.

Güthe, S., Kapinos, L., Möglich, A., Meier, S., Grzesiek, S., and Kiefhaber, T. (2004), "Very fast folding and association of a trimerization domain from bacteriophage T4 fibritin," *Journal of molecular biology*, 337, 905–915.

Halperin, D. (1997), "Arrangements," in *Handbook of discrete and computational geometry*, eds. J. E. Goodman and J. O'Rourke, pp. 529–562, CRC Press, Inc., Boca Raton, FL, USA, Second edn.

Hanniel, I. and Halperin, D. (2000), "Two-dimensional arrangements in CGAL and Adaptive Point Location for Parametric Curves," in *Proc. 4th Workshop on Algorithm Engineering*, vol. 1982 of *Lecture Notes in Computer Science*, pp. 171–182, Springer-Verlag.

Harrison, S. C. (2008), "Viral membrane fusion," *Nature structural & molecular biology*, 15, 690–698.

Haynes, B. F., Fleming, J., Clair, E. W. S., Katinger, H., Stiegler, G., Kunert, R., Robinson, J., Scearce, R. M., Plonk, K., Staats, H. F., et al. (2005), "Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies," *Science*, 308, 1906–1908.

Herrmann, T., Güntert, P., and Wüthrich, K. (2002), "Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA," *Journal of Molecular Biology*, 319, 209–227.

Hinz, A., Schoehn, G., Quendler, H., Hulsik, D. L., Stiegler, G., Katinger, H., Seaman, M. S., Montefiori, D., and Weissenhorn, W. (2009), "Characterization of a trimeric MPER containing HIV-1 gp41 antigen," *Virology*, 390, 221–227.

Huang, J., Ofek, G., Laub, L., Louder, M. K., Doria-Rose, N. A., Longo, N. S., Imamichi, H., Bailer, R. T., Chakrabarti, B., Sharma, S. K., et al. (2012), "Broad and potent neutralization of HIV-1 by a gp41-specific human antibody," *Nature*, 491, 406–412.

Hungerford, T. (1980), *Algebra*, Springer, New York.

Hus, J. C., Marion, D., and Blackledge, M. (2001), "Determination of protein backbone structure using only residual dipolar couplings." *J. Am. Chem. Soc.*, 123, 1541–2.

Ikura, M. and Bax, A. (1992), "Isotope-filtered 2D NMR of a protein-peptide complex: study of a skeletal muscle myosin light chain kinase fragment bound to calmodulin," *Journal of the American Chemical Society*, 114, 2433–2440.

Kay, L. E., Clore, G. M., Bax, A., and Gronenborn, A. M. (1990), "Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1 beta in solution," *Science*, 249, 411–414.

Kim, S. and Szyperski, T. (2003), "GFT NMR, a New Approach To Rapidly Obtain Precise High-Dimensional NMR Spectral Information," *Journal of the American Chemical Society*, 125, 1385–1393.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992), "Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor." *Science*, 256, 1783–90.

Kovacs, H., O'Donoghue, S. I., Hoppe, H.-J., Comfort, D., Reid, K. B. M., Campbell, I. D., and Nilges, M. (2002), "Solution structure of the coiled-coil trimerization domain from lung surfactant protein D," *Journal of Biomolecular NMR*, 24, 89–102, 10.1023/A:1020980006628.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (1998), "Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody." *Nature*, 393, 648–59.

Lange, O. F., Lakomek, N.-A., Farès, C., Schröder, G. F., Walter, K. F. A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C., and de Groot, B. L. (2008), "Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution." *Science*, 320, 1471–5.

Langmead, C. J., Yan, A., Lilien, R., Wang, L., and Donald, B. R. (2004), "A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments." *J. Comput. Biol.*, 11, 277–98.

Lenz, O., Dittmar, M. T., Wagner, A., Ferko, B., Vorauer-Uhl, K., Stiegler, G., and Weissenhorn, W. (2005), "Trimeric membrane-anchored gp41 inhibits HIV membrane fusion," *Journal of Biological Chemistry*, 280, 4095–4101.

Li, D., Lyons, J. A., Pye, V. E., Vogeley, L., Aragão, D., Kenyon, C. P., Shah, S. T. A., Doherty, C., Aherne, M., and Caffrey, M. (2013), "Crystal structure of the integral membrane diacylglycerol kinase." *Nature*, 497, 521–4.

Linge, J. P., O'Donoghue, S. I., and Nilges, M. (2001), "Automated assignment of ambiguous nuclear overhauser effects with ARIA." *Meth. Enzymol.*, 339, 71–90.

Linge, J. P., Habeck, M., Rieping, W., and Nilges, M. (2004), "Correction of spin diffusion during iterative automated NOE assignment," *Journal of Magnetic Resonance*, 167, 334–342.

Lipfert, J., Columbus, L., Chu, V. B., Lesley, S. A., and Doniach, S. (2007), "Size and shape of detergent micelles determined by small-angle X-ray scattering," *The Journal of Physical Chemistry B*, 111, 12427–12438.

Liu, J., Deng, Y., Dey, A. K., Moore, J. P., and Lu, M. (2009), "Structure of the HIV-1 gp41 Membrane-Proximal Ectodomain Region in a Putative Prefusion Conformation," *Biochemistry*, 48, 2915–2923.

Long, D. and Brüschweiler, R. (2011), "In silico elucidation of the recognition dynamics of ubiquitin." *PLoS Comput. Biol.*, 7, e1002035.

Losonczi, J. A., Andrec, M., Fischer, M. W. F., and Prestegard, J. H. (1999), "Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition," *Journal of Magnetic Resonance*, 138, 334–342.

Lozano-Perez, T. (1981), "Automatic Planning of Manipulator Transfer Movements," *Systems, Man and Cybernetics, IEEE Transactions on*, 11, 681–698.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.-H., Haim, H., Yang, X., and Sodroski, J. (2012), "Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer," *Nature structural & molecular biology*.

Margolis, D. M. and Archin, N. M. (2006), "Attacking HIV provirus: therapeutic strategies to disrupt persistent infection." *Infect Disord Drug Targets*, 6, 369–76.

Marion, D., Kay, L. E., Sparks, S. W., Torchia, D. A., and Bax, A. (1989), "Three-dimensional heteronuclear NMR of nitrogen-15 labeled proteins," *Journal of the American Chemical Society*, 111, 1515–1517.

Martin, J. W., Yan, A. K., Bailey-Kellogg, C., Zhou, P., and Donald, B. R. (2011a), "A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs," *Journal of Computational Biology*, 18, 1507–1523.

Martin, J. W., Yan, A. K., Bailey-Kellogg, C., Zhou, P., and Donald, B. R. (2011b), "A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs," in *Proceedings of The Fifteenth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, eds. V. Bafna and S. C. Sahinalp, pp. 222–237, Vancouver, BC, Springer Berlin, RECOMB 2011, Lecture Notes in Computer Science, LNBI 6577.

Martin, J. W., Yan, A. K., Bailey-Kellogg, C., Zhou, P., and Donald, B. R. (2011c), "A graphical method for analyzing distance restraints using residual dipolar couplings for structure determination of symmetric protein homo-oligomers," *Protein Science*, 20, 970–985.

McCune, J. M., Rabin, L. B., Feinberg, M. B., Lieberman, M., Kosek, J. C., Reyes, G. R., and Weissman, I. L. (1988), "Endoproteolytic cleavage of gp160 is required for the activation of human immunodeficiency virus." *Cell*, 53, 55–67.

Muñoz-Barroso, I., Salzwedel, K., Hunter, E., and Blumenthal, R. (1999), "Role of the membrane-proximal domain in the initial stages of human immunodeficiency virus type 1 envelope glycoprotein-mediated membrane fusion," *Journal of virology*, 73, 6089–6092.

Muster, T., Steindl, F., Purtscher, M., Trkola, A., Klima, A., Himmler, G., Rüker, F., and Katinger, H. (1993), "A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1." *Journal of Virology*, 67, 6642–6647.

Nilges, M. (1993), "A calculation strategy for the structure determination of symmetric dimers by ^{1}H NMR," *Proteins: Structure, Function, and Genetics*, 17, 297–309.

Nilges, M. and O'Donoghue, S. I. (1998), "Ambiguous NOEs and automated NOE assignment," *Progress in Nuclear Magnetic Resonance Spectroscopy*, 32, 107–139.

Nilges, M., Malliavin, T., and Bardiaux, B. (2010), "Protein structure calculation using ambiguous restraints," *eMagRes*.

O'Donoghue, S. I., Chang, X., Abseher, R., Nilges, M., and Led, J. J. (2000), "Unraveling the symmetry ambiguity in a hexamer: Calculation of the R6 human insulin structure," *Journal of Biomolecular NMR*, 16, 93–108, 10.1023/A:1008323819099.

Ofek, G., Tang, M., Sambor, A., Katinger, H., Mascola, J. R., Wyatt, R., and Kwong, P. D. (2004), "Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope," *Journal of virology*, 78, 10724–10737.

Ofek, G., Guenaga, F. J., Schief, W. R., Skinner, J., Baker, D., Wyatt, R., and Kwong, P. D. (2010a), "Elicitation of structure-specific antibodies by epitope scaffolds," *Proceedings of the National Academy of Sciences*, 107, 17880–17887.

Ofek, G., McKee, K., Yang, Y., Yang, Z.-Y., Skinner, J., Guenaga, F. J., Wyatt, R., Zwick, M. B., Nabel, G. J., Mascola, J. R., et al. (2010b), "Relationship between antibody 2F5 neutralization of HIV-1 and hydrophobicity of its heavy chain third complementarity-determining region," *Journal of virology*, 84, 2955–2962.

Oxenoid, K. and Chou, J. J. (2005), "The structure of phospholamban pentamer reveals a channel-like architecture in membranes," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 10870–10875.

Ozkan, E., Yu, H., and Deisenhofer, J. (2005), "Mechanistic insight into the allosteric activation of a ubiquitin-conjugating enzyme by RING-type ubiquitin ligases." *Proc. Natl. Acad. Sci. U.S.A.*, 102, 18890–5.

Phogat, S., Svehla, K., Tang, M., Spadaccini, A., Muller, J., Mascola, J., Berkower, I., and Wyatt, R. (2008), "Analysis of the human immunodeficiency virus type 1 gp41 membrane proximal external region arrayed on hepatitis B surface antigen particles," *Virology*, 373, 72–84.

Potluri, S., Yan, A. K., Chou, J. J., Donald, B. R., and Bailey-Kellogg, C. (2006), "Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space, using NMR restraints and van der Waals packing," *Proteins: Structure, Function, and Bioinformatics*, 65, 203–219.

Potluri, S., Yan, A. K., Donald, B. R., and Bailey-Kellogg, C. (2007), "A complete algorithm to resolve ambiguity for intersubunit NOE assignment in structure determination of symmetric homo-oligomers," *Protein Science*, 16, 69–81.

Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T. E., and Nilges, M. (2007), "ARIA2: Automated NOE assignment and data integration in NMR structure calculation," *Bioinformatics*, 23, 381–382.

Salmon, L., Bascom, G., Andricioaei, I., and Al-Hashimi, H. M. (2013), "A General Method for Constructing Atomic-Resolution RNA Ensembles using NMR Residual Dipolar Couplings: The Basis for Interhelical Motions Revealed," *Journal of the American Chemical Society*, 135, 5457–5466.

Salzwedel, K., West, J. T., and Hunter, E. (1999), "A conserved tryptophan-rich motif in the membrane-proximal region of the human immunodeficiency virus type 1 gp41 ectodomain is important for Env-mediated fusion and virus infectivity," *Journal of virology*, 73, 2469–2480.

Sato, S., Religa, T. L., Daggett, V., and Fersht, A. R. (2004), "Testing protein-folding simulations by experiment: B domain of protein A," *Proc. Natl. Acad. Sci. U.S.A.*, 101, 6952–6956.

Scherer, E. M., Leaman, D. P., Zwick, M. B., McMichael, A. J., and Burton, D. R. (2010), "Aromatic residues at the edge of the antibody combining site facilitate viral glycoprotein recognition through membrane interactions," *Proceedings of the National Academy of Sciences*, 107, 1529–1534.

Schibli, D. J., Montelaro, R. C., and Vogel, H. J. (2001), "The membrane-proximal tryptophan-rich region of the HIV glycoprotein, gp41, forms a well-defined helix in dodecylphosphocholine micelles," *Biochemistry*, 40, 9570–9578.

Schnell, J. R. and Chou, J. J. (2008), "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature*, 451, 591–595, 10.1038/nature06531.

Schrödinger L.L.C. (2012), "The PyMOL Molecular Graphics System," Version 1.5.0.1.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003), "The Xplor-NIH NMR molecular structure determination package," *Journal of Magnetic Resonance*, 160, 65–73.

Schwieters, C. D., Kuszewski, J. J., and Marius Clore, G. (2006), "Using Xplor-NIH for NMR molecular structure determination," *Progress in Nuclear Magnetic Resonance Spectroscopy*, 48, 47–62, doi: DOI: 10.1016/j.pnmrs.2005.10.001.

Shahied, L., Braswell, E. H., LeStourgeon, W. M., and Krezel, A. M. (2001), "An antiparallel four-helix bundle orients the high-affinity RNA binding sites in hnRNP C: a mechanism for RNA chaperonin activity." *J. Mol. Biol.*, 305, 817–28.

Si, Z., Madani, N., Cox, J. M., Chruma, J. J., Klein, J. C., Schön, A., Phan, N., Wang, L., Biorn, A. C., Cocklin, S., Chaiken, I., Freire, E., Smith, A. B., and Sodroski, J. G. (2004), "Small-molecule inhibitors of HIV-1 entry block receptor-induced conformational changes in the viral envelope glycoproteins." *Proc. Natl. Acad. Sci. U.S.A.*, 101, 5036–41.

Song, L., Sun, Z.-Y. J., Coleman, K. E., Zwick, M. B., Gach, J. S., Wang, J.-h., Reinherz, E. L., Wagner, G., and Kim, M. (2009), "Broadly neutralizing anti-HIV-1 antibodies disrupt a hinge-related function of gp41 at the membrane interface." *Proc. Natl. Acad. Sci. U.S.A.*, 106, 9057–62.

Stachelhaus, T. and Marahiel, M. A. (1995), "Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA." *J. Biol. Chem.*, 270, 6163–9.

Stiegler, G., Kunert, R., Purtscher, M., Wolbank, S., Voglauer, R., Steindl, F., and Katinger, H. (2001), "A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1," *AIDS research and human retroviruses*, 17, 1757–1765.

Sun, Z.-Y. J., Oh, K. J., Kim, M., Yu, J., Brusic, V., Song, L., Qiao, Z., Wang, J.-h., Wagner, G., and Reinherz, E. L. (2008), "HIV-1 broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane," *Immunity*, 28, 52–63.

Tao, Y., Strelkov, S. V., Mesyanzhinov, V. V., and Rossmann, M. G. (1997), "Structure of bacteriophage T4 fibritin: a segmented coiled coil and the role of the C-terminal domain." *Structure*, 5, 789–98.

Taylor, W. R., Bartlett, G. J., Chelliah, V., Klose, D., Lin, K., Sheldon, T., and Jonassen, I. (2008), "Prediction of protein structure from ideal forms." *Proteins*, 70, 1610–9.

Tilton, J. C. and Doms, R. W. (2010), "Entry inhibitors in the treatment of HIV-1 infection," *Antiviral Research*, 85, 91–100.

Tolman, J. R. (2002), "A novel approach to the retrieval of structural and dynamic information from residual dipolar couplings using several oriented media in biomolecular NMR spectroscopy," *Journal of the American Chemical Society*, 124, 12020–12030.

Tolman, J. R. and Ruan, K. (2006), "NMR residual dipolar couplings as probes of biomolecular dynamics." *Chem. Rev.*, 106, 1720–36.

Tran, E. E. H., Borgnia, M. J., Kuybeda, O., Schauder, D. M., Bartesaghi, A., Frank, G. A., Sapiro, G., Milne, J. L. S., and Subramaniam, S. (2012), "Structural mechanism of trimeric HIV-1 envelope glycoprotein activation." *PLoS Pathog.*, 8, e1002797.

Tripathy, C., Zeng, J., Zhou, P., and Donald, B. R. (2011), "Protein loop closure using orientational restraints from NMR data." *Proteins*.

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2007), "BioMagResBank," *Nucl. Acids Res.*, pp. D402–D408.

Van Horn, W. D., Kim, H.-J., Ellis, C. D., Hadziselimovic, A., Sulistijo, E. S., Karra, M. D., Tian, C., Sönnichsen, F. D., and Sanders, C. R. (2009), "Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase." *Science*, 324, 1726–9.

Verkoczy, L., Diaz, M., Holl, T. M., Ouyang, Y.-B., Bouton-Verville, H., Alam, S. M., Liao, H.-X., Kelsoe, G., and Haynes, B. F. (2010), "Autoreactivity in an HIV-1 broadly reactive neutralizing antibody variable region heavy chain induces immunologic tolerance," *Proceedings of the National Academy of Sciences*, 107, 181–186.

Verkoczy, L., Chen, Y., Zhang, J., Bouton-Verville, H., Newman, A., Lockwood, B., Scearce, R. M., Montefiori, D. C., Dennison, S. M., Xia, S.-M., et al. (2013), "Induction of HIV-1 Broad Neutralizing Antibodies in 2F5 Knock-in Mice: Selection against Membrane Proximal External Region–Associated Autoreactivity Limits T-Dependent Responses," *The Journal of Immunology*, 191, 2538–2550.

Vinogradova, O., Sönnichsen, F., and Sanders, C. R. (1998), "On choosing a detergent for solution NMR studies of membrane proteins." *J. Biomol. NMR*, 11, 381–6.

Wang, C.-s. E., Lozano-Pérez, T., and Tidor, B. (1998), "AmbiPack: A systematic algorithm for packing of macromolecular structures with ambiguous distance constraints," *Proteins: Structure, Function, and Genetics*, 32, 26–42.

Wang, J., Pielak, R. M., McClintock, M. A., and Chou, J. J. (2009), "Solution structure and functional analysis of the influenza B proton channel," *Nat Struct Mol Biol*, 16, 1267–1271, 10.1038/nsmb.1707.

Wang, L. and Donald, B. R. (2004), "Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure," *Journal of Biomolecular NMR*, 29, 223–242.

Wang, X., Bansal, S., Jiang, M., and Prestegard, J. H. (2008), "RDC-assisted modeling of symmetric protein homo-oligomers," *Protein Science*, 17, 899–907.

Wedemeyer, W. J., Rohl, C. A., and Scheraga, H. A. (2002), "Exact solutions for chemical bond orientations from residual dipolar couplings," *Journal of Biomolecular NMR*, 22, 137–151, 10.1023/A:1014206617752.

Welzl, E. (1991), *Smallest enclosing disks (balls and ellipsoids)*, Springer.

Werner-Allen, J. W., Coggins, B. E., and Zhou, P. (2010), "Fast acquisition of high resolution 4-D amide-amide NOESY with diagonal suppression, sparse sampling and FFT-CLEAN." *J. Magn. Reson.*, 204, 173–8.

White, S. H. (2004), "The progress of membrane protein structure determination," *Protein Science*, 13, 1948–1949.

White, T. A., Bartesaghi, A., Borgnia, M. J., Meyerson, J. R., de la Cruz, M. J. V., Bess, J. W., Nandwani, R., Hoxie, J. A., Lifson, J. D., Milne, J. L., et al. (2010), "Molecular architectures of trimeric SIV and HIV-1 envelope glycoproteins on intact viruses: strain-dependent variation in quaternary structure," *PLoS pathogens*, 6, e1001249.

Wu, S.-R., Löving, R., Lindqvist, B., Hebert, H., Koeck, P. J., Sjöberg, M., and Garoff, H. (2010), "Single-particle cryoelectron microscopy analysis reveals the HIV-1 spike as a tripod structure," *Proceedings of the National Academy of Sciences*, 107, 18844–18849.

Wüthrich, K., Billeter, M., and Braun, W. (1983), "Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance," *Journal of Molecular Biology*, 169, 949–961.

Yang, G., Holl, T. M., Liu, Y., Li, Y., Lu, X., Nicely, N. I., Kepler, T. B., Alam, S. M., Liao, H.-X., Cain, D. W., et al. (2013), "Identification of autoantigens recognized by the 2F5 and 4E10 broadly neutralizing HIV-1 antibodies," *The Journal of experimental medicine*, 210, 241–256.

Yershova, A., Jain, S., Lavalle, S. M., and Mitchell, J. C. (2010), "Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration." *Int J Rob Res*, 29, 801–812.

Zanetti, G., Briggs, J. A., Grünewald, K., Sattentau, Q. J., and Fuller, S. D. (2006), "Cryo-electron tomographic structure of an immunodeficiency virus envelope complex in situ," *PLoS pathogens*, 2, e83.

Zeng, J., Boyles, J., Tripathy, C., Wang, L., Yan, A., Zhou, P., and Donald, B. (2009), "High-resolution protein structure determination starting with a global fold calculated from exact solutions to the RDC equations," *Journal of Biomolecular NMR*, 45, 265–281.

Zeng, J., Zhou, P., and Donald, B. R. (2011), "Protein side-chain resonance assignment and NOE assignment using RDC-defined backbones without TOCSY data." *J. Biomol. NMR*, 50, 371–95.

Zheng, D., Aramini, J. M., and Montelione, G. T. (2004), "Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data." *Protein Sci.*, 13, 549–54.

Zhu, P., Liu, J., Bess, J., Chertova, E., Lifson, J. D., Grisé, H., Ofek, G. A., Taylor, K. A., and Roux, K. H. (2006), "Distribution and three-dimensional structure of AIDS virus envelope spikes," *Nature*, 441, 847–852.

Zwick, M. B., Labrijn, A. F., Wang, M., Spenlehauer, C., Saphire, E. O., Binley, J. M., Moore, J. P., Stiegler, G., Katinger, H., Burton, D. R., et al. (2001), "Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41," *Journal of virology*, 75, 10892–10905.

Biography

Jeffrey W. Martin was born on March 11, 1983 in Florissant, MO to Joanne and Edward Martin. He received his Bachelor of Science in computer science at the University of Missouri – Rolla in 2004. UMR has since been renamed to the Missouri University of Science and Technology. In 2009, Jeff received his Master of Science in computer science from Duke University. He joined the doctoral program in computer science at Duke University in 2009.

Jeff received the Oustanding Master's Thesis Award in 2009 for his Master's thesis, *On algorithms for structure determination of symmetric proteins from nuclear magnetic resonance data*. In 2010, Jeff received the Outstanding Departmental Service award for his efforts to improve the North Building office space for the incoming graduate students.

List of publications

"Structure of an HIV-1 Neutralizing Antibody Target: A Lipid Bound gp41 Envelope Membrane Proximal Region Trimer" (with Patrick N. Reardon, Harvey Sage, S. Moses Dennison, S. Munir Alam, Barton F. Haynes, Bruce R. Donald, and Leonard D. Spicer) *Proceedings of the National Academy of Sciences* (in press), 2013.

"A graphical method for analyzing distance restraints using residual dipolar couplings for structure determination of symmetric protein homo-oligomers" (with

Anthony K. Yan, Pei Zhou, Chris Bailey-Kellogg, and Bruce R. Donald) *Journal of Protein Science*, 20(6): 970-985, 2011.

"A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs" (with Anthony K. Yan, Pei Zhou, Chris Bailey-Kellogg, and Bruce R. Donald) *Journal of Computational Biology*, 18(11): 1507-1523, 2011.

"A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs" (with Anthony K. Yan, Pei Zhou, Chris Bailey-Kellogg, and Bruce R. Donald) *Proceedings of The Fifteenth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* Vancouver, BC, 2011.

"Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints" (with Bruce R. Donald) *Progress in NMR Spectroscopy* 55(2):101127, 2009.